

**U.S. Tiwary  
Tanveer J. Siddiqui  
M. Radhakrishna  
M. D. Tiwari  
(Editors)**

**Proceeding of the  
First International Conference on  
Intelligent Human  
Computer Interaction  
(IHCI 2009)**



U. S. Tiwary · Tanveer J. Siddiqui ·  
M. Radhakrishna · M. D. Tiwari

Proceedings of the First  
International Conference on  
**INTELLIGENT  
HUMAN COMPUTER  
INTERACTION**

(IHCI 2009) January 20–23, 2009

Organized by the Indian Institute of Information  
Technology, Allahabad, India



IHCI 2009

 Springer

**First International Conference  
on  
Intelligent Human Computer  
Interaction  
(IHCI 2009)**

Januaury 20–23, 2009

Organized By

Indian Institute of Information Technology,  
Allahabad, India

&

IEEE Workshop on Recent Trends in Human Computer  
Interaction  
January 19–21, 2009

Organized by

IEEE UP Chapter, IIT Kanpur  
&

Indian Institute of Information Technology, Allahabad



**Editors**

U. S. Tiwary  
Tanveer J. Siddiqui  
M. Radhakrishna  
M. D. Tiwari

**Editors**

**U.S. Tewary**  
Professor, Indian Institute of Information  
Information  
Technology  
Allahabad, India

**Tanveer J. Siddiqui**  
Assistant Professor,  
Indian Institute of Information  
Technology  
Allahabad, India

**M.Radhakrishna**  
Professor, Indian Institute of  
Technology  
Allahabad, India

**M.D Tewari**  
Director,  
Indian Institute of Information  
Technology  
Allahabad, India

ISBN: 978-81-8489-203-1

© 2009 Indian Institute of Information Technology, India  
Deoghat, Jhalwa  
Allahabad-211012  
India

All rights reserved. No part of the book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without written permission from the copyright holder and the publisher, except for brief excerpts in connection with reviews or scholarly analysis. The use in this publication of trademarks, radenames, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

All comments, opinions, conclusions, or recommendations in the articles are those of the author(s), and do not necessarily reflect the views of the publisher.

Published by Springer (India) Private Limited;  
Typeset, Edited and Proofread by IHCI 2009 team and  
Printed by Rajkamal Electric Press, Delhi- 110 033.

This is a special conference proceedings published in continuation with the First International Conference on Intelligent human computer Interaction (IHCI 2009) Organized by the Indian Institute of information Technology during 20 January 2009 through 23 January 2009.

**Dedicated to**

***Charles Darwin***

*At his bi-centenary year who revealed the  
fact that “humanity is our own creation”*

## **Program Committee**

### **Chief Patron**

F C Kohli

### **Patron**

M D Tiwari

### **General Chair**

Mriganka Sur

### **Co-Chairs**

U. S. Tiwary

M. Radhakrishna

### **Coordinator**

Tanveer J. Siddiqui

## **International Advisory Committee**

Alexander Gelbukh, Mexico	Keith Cheverst, UK
Alexander Pasko, UK	L M Patnaik, India
Amanda Spink, Australia	L V Subramaniam, India
Arvind Joshi, USA	M. M. Sufyan Beg, India
Dominik Heckmann, Germany	Manuel Montes Gomez, Mexico
Eyas-Al-Qawasmeh, Jordan	Michael Gamon, USA
Francisco J. Serón, Spain	N Balakrishnan, India
Fuji Ren, Japan	Pawan Sinha, USA
Gabriella Pasi, Italy	Pushpak Bhattacharya, India
Geehyuk Lee, South Korea	R M K Sinha, India
Hyo-Sung Ahn, South Korea	Rada Mihalcea, USA
Hyo-Sung Ahn, South Korea	Richard Chebeir, France
Iadh Ounis, Glasgow UK	S K Tripathi, USA
Iryna Gurevych, Germany	Vaclav Snasel, Czech Republic
James F. Peters, Canada	Youakim Badr, France

## **Review Committee**

Alexander Gelbukh, AMC, Mexico  
Alexander Pasko, Bournemouth University, UK  
Amanda Spink, Queensland University, Australia  
Anupam Basu, IIT Kharagpur, India  
Arvind Joshi, University of Pennsylvania, USA  
B B Choudhury, ISI Kolkata, India  
Dominik Heckmann, Saarland University, Germany  
Eyas-Al-Qawasmeh, JUST, Jordan  
Francisco J. Serón, University of Zaragoza, Spain  
Fuji Ren, The University of Tokushima, Japan  
Gabriella Pasi, University of Milano, Bicocca, Italy  
Gael Dias, University of Beira Interior, Portugal  
Gaël Harry Dias, Portugal  
Galina Pasko, European University of Lefke, TRNC  
Geehyuk Lee, ICU, South Korea  
Hadeel Nasrat Abdullah, University of Technology, Bagdad, Iraq  
Hammadi Nait-Charif, Bournemouth University, UK  
Hyo-Sung Ahn, GIST, South Korea  
Hyo-Sung Ahn, South Korea  
Iadh Ounis, Glasgow University, Glasgow, UK  
Ilham Huseynov, European University of Lefke, TRNC  
Iryna Gurevych, TU Darmstadt, Germany  
James F. Peters, University of Manitoba, Canada  
Keith Cheverst, Lancaster University, UK  
L M Patnaik, IISC Bangalore, India  
L V Subramaniam, IBM India Research Lab, India  
M. M. Sufyan Beg, JMI, India  
Mahua Bhattacharya, IIITM Gwalior, India  
Manuel Montes Gomez, INAOE, Mexico  
Michael Gamon, Microsoft, USA  
N Balakrishnan, IISc, Bangalore, India  
Partha Bhowmick, IIT, Kharagpur, India  
Pawan Sinha, MIT, USA  
Pushpak Bhattacharya, IITB, India  
R M K Sinha, IITK, India  
Rada Mihalcea, UNT, USA  
Richard Chebeir, France  
S K Tripathi, SUNY, USA  
Shekahr Verma, IIIT Allahabad, India  
Shirshu Verma, IIIT Allahabad, India  
Sumana Gupta, IIT Kanpur, India  
Tanveer J Siddiqui, IIIT Allahabad, India  
Tapobrat Lahiri, IIIT Allahabad, India  
UmaShanker Tiwary, IIIT Allahabad, India  
Vaclav Snasel, TU of Ostrava, Czech Republic  
Youakim Badr, INSA de Lyon, France

# Preface

Dear Reader!

Welcome to the proceedings of the First International Conference on Intelligent Human Computer Interaction (IHCI 2009) organized by the Indian Institute of Information Technology Allahabad. This is the first International Conference focused on Human Computer Interaction being organized in India. There is an increased interest in the human factors issues of computer use with a number of systems. The conference aims to provide an excellent opportunity for the dissemination of interesting new research, discussion about them and the generation of new ideas in these areas.

We planned to organize the conference around the following five tracks:

- Signal and Vision Processing
- Language Processing
- Cognitive modeling
- Sensors and Embedded systems for HCI
- Graphics, Animation and Gaming

Graphics, Animation and Gaming, Signal and Vision Processing, Language Processing and Cognitive modeling attracted due attention in the conference program. Very few papers were submitted in Sensors and Embedded systems and Graphics and Animation. Language is the primary means of communication between humans though usually neglected from HCI perspective. It will be better if human-computer interaction can be done in the same model as human-human communication. This was the main motivation behind including Language Processing as a separate track in the conference. However, some of the papers could not really achieve the application aspect from the HCI perspective. We will improve on this point in the next conference.

In total nearly 130 papers were submitted. The program committee had a very challenging task of selecting high quality submissions. Each paper was peer-reviewed by at least two reviewers. On the recommendation of Program Committee 41 papers were accepted for the presentation in the main conference including two papers accepted for Web X.0 and Web Mining workshop. Out of which six papers were withdrawn. Remaining thirty five papers put were re-organized in following four sessions:

- Signal and Vision Processing
- Human Computer Interaction
- Language Processing

### Knowledge Discovery and Data Mining

Thirty five papers and four talks put in the proceedings of the First International Conference on Intelligent Human Computer Interaction cover a wide spectrum of research topics in HCI like brain-computer interface, text and vision based interfaces, visualization tools etc.

This volume will be useful for researchers working in this fascinating and fast growing research area.

We would like to express our sincere thanks to all the invited speakers, tutorial presenters, authors and members of the program committee that has made this conference a success.

U. S. Tiwary

Tanveer J. Siddiqui

M Radhakrishna

M. D. Tiwari

## **Acknowledgement**

We sincerely thank to our honourable patrons Sri F. C. Kohli and Dr. M. D. Tiwari for their whole-hearted support for this conference.

We thank our conference chair, Prof. Mriganka Sur, from deepest corner of our heart.

This conference would not have been successful without the support of IIIT-Allahabad. Our sincere thank goes to the entire family of IIIT-Allahabad, including faculty, staff and students for their assistance and hard work in organizing the conference and its proceedings.

We gratefully acknowledge IEEE UP chapter, IIT Kanpur for co-sponsoring the workshop ‘Recent trends in HCI’.

We also thank Patent Referral Centre, IIIT Allahabad, setup by Ministry of Communication and Information Technology, Government of India, for conducting the plagiarism check on all the technical papers submitted.

And last but not the least, we express our deep sense of gratitude to all the invited speakers, tutorial presenters, authors and members of the program committee without their support this conference could not have been successful.

U. S. Tiwary & M. Radhakrishna  
Co-Chairs,  
IHCI 2009

# Table of Contents

## Invited Lectures

Exploring the Intersection of HCI and Ubicomp: The Design, Deployment and Use of Situated Displays .....	3
<i>Keith Cheverst</i>	
Evolution of Geometric Figures from the Euclidean to Digital Era .....	19
<i>Partha Bhowmick</i>	
Web Content Mining Focused on Named Objects .....	37
<i>Václav Snášel and Milos Kudelka</i>	
Words and Pictures: An HCI Perspective .....	59
<i>Tanveer J. Siddiqui and Uma Shanker Tiwary</i>	

## Signal and Vision Processing

An Application for Driver Drowsiness Identification based on Pupil Detection using IR Camera .....	73
<i>K. S. Chidanand Kumar and Brojeshwar Bhowmick</i>	
Engine Fault Diagnosis Using DTW, MFCC and FFT .....	83
<i>Vrijendra Singh and Narendra Meena</i>	
Multimodal News Story Segmentation .....	95
<i>Gert-Jan Poulsse and Marie-Francine Moens</i>	
RGB Color Histogram Feature based Image Classification: An Application of Rough Reasoning .....	102
<i>Shailendra Singh</i>	

- A Robust Object Tracking Method for Noisy Video  
using Rough Entropy in Wavelet Domain ..... 113  
*Anand Singh Jalal and Uma Shanker Tiwary*

## Human Computer Interaction

- Extraction of Rhythmic Information from Non-Invasively  
Recorded EEG Signal Using  
IEEE Standard 1057 Algorithm ..... 125  
*Manoj Kumar Mukul and Fumitoshi Matsuno*
- Registration of Multimodality Medical Imaging of  
Brain using Particle Swarm Optimization ..... 131  
*Mahua Bhattacharya and Arpita Das*
- Relative Amplitude based Feature of Characteristic ECG Peaks for  
Identification of Cornary Artory Disease ..... 140  
*Bakul Gohel, U. S. Tiwary and T. Lahiri*
- ProVis: An Anaglyph based Visualization Tool for  
Protein Molecules ..... 147  
*Rajesh Bhasin and Abhishek Kumar*
- Applying Cognitive Psychology to User Interfaces ..... 156  
*Sabeen Durrani and Qaiser S. Durrani*
- CAST: A Novel Trajectory Clustering and Visualization tool for  
Spatio-Temporal Data ..... 169  
*Hazarath Munaga, Lucio Ieronutti and Luca Chittaro*
- Vote Stuffing Control in IPTV-based Recommender  
Systems ..... 176  
*Rajen Bhatt*
- Static and Dynamic Features for Improved HMM based  
Visual Speech Recognition ..... 184  
*R. Rajavel and P. S. Sathidevi*

A Single Accelerometer based Wireless Embedded System for Predefined Dynamic Gesture Recognition . . . . .	195
<i>Rahul Parsani and Karandeep Singh</i>	
What Combinations of Contents is Driving Popularity in IPTV-based Social Networks? . . . . .	202
<i>Rajen Bhatt</i>	
Adaptive Acceleration of MAP with Entropy Prior and Flux Conservation for Image Deblurring . . . . .	212
<i>Manoj Kumar Singh, Yong-Hoon Kim, U. S. Tiwary, Rajkishore Prasad and Tanveer Siddiqui</i>	

## Language Processing

An Intelligent Automatic Text Summarizer . . . . .	223
<i>M. Shoaib Jameel, Anubhav, Nilesh Singh, Nitin Kumar Singh, Chingtham Tejbanta Singh and M. K. Ghose</i>	
A Hybrid Approach for Transliteration of Name Entities . . . . .	231
<i>R. C. Balabantaray, S. Mohanty and R. K. Das</i>	
Information Uptriever: A system for Content Assimilation and Aggregation for Developing Regions . . . . .	241
<i>Ravindra Dastikop, G. A. Radde and Jaydevi C. Karur</i>	
Classification of Products through Blog Analysis . . . . .	246
<i>Niladri Chatterjee, Sumit Bisai and Prasenjit Chakraborty</i>	
Document Summarization using Wikipedia . . . . .	254
<i>Krishnan Ramanathan, Yogesh Sankarasubramaniam, Nidhi Mathur and Ajay Gupta</i>	
Improving Performance of English–Hindi CLIR System using Linguistic Tools and Techniques . . . . .	261
<i>Anurag Seetha, Sujoy Das and M. Kumar</i>	

Improving Multi-document Text Summarization Performance using Local and Global Trimming .....	272
<i>Kamal Sarkar</i>	
STAIR : A System for Topical and Aggregated Information Retrieval .....	283
<i>C. V. Krishnakumar and Krishnan Ramanathan</i>	
Exploring Multiple Ontologies and WordNet Framework to Expand Query for Question Answering System .....	296
<i>Santosh Kumar Ray, Shailendra Singh and B. P. Joshi</i>	
Disambiguation Strategies for English–Hindi Cross Language Information Retrieval System .....	306
<i>Sujoy Das, Anurag Seetha, M. Kumar and J. L. Rana</i>	
Evaluating Effect of Stemming and Stop-word Removal on Hindi Text Retrieval .....	316
<i>Amaresh K Pandey, Tanveer J Siddiqui</i>	
An Unsupervised Approach to Hindi Word Sense Disambiguation .....	327
<i>Neetu Mishra, Shashi Yadav and Tanveer J. Siddiqui</i>	
Search Result Clustering with using a Singular Value Decomposition (SVD) .....	336
<i>Hussam Dahwa Abdulla and Vaclav Snasel</i>	
Automatic Performance Evaluation of Web Search Systems using Rough Set Based Rank Aggregation .....	344
<i>Rashid Ali and M. M. Sufyan Beg</i>	

## **Knowledge Discovery and Data Mining**

Minimizing Space-Time Complexity in RSTBD a New Method for Frequent Pattern Mining .....	361
<i>Vaibhav Kant Singh and Vinay Kumar Singh</i>	

Privacy Preserving Data Mining: A New Methodology for Data Transformation .....	372
<i>A. K. Upadhyay, Abhijat Agarwal, Rachita Masand     and Rajeev Gupta</i>	
Data Aggregation in Cluster based Wireless Sensor Networks .....	391
<i>Shirshu Verma and Uma Shanker Tiwary</i>	

# Invited Lectures

# Exploring the Intersection of HCI and Ubicomp: The Design, Deployment and use of Situated Displays

Keith Cheverst

Infolab21, Lancaster University  
kc@comp.lancs.ac.uk

**Abstract.** In this chapter I describe my exploration of the design, deployment and use of Situated Displays. I argue and illustrate how research on Situated displays provides a useful vehicle for exploring issues that arise at the intersection of ubicomp and HCI. Following on from Weiser's early definition of ubicomp, I focus on the importance of understanding settings for ubicomp deployments in order to enable the successful 'weaving in' of technologies into the existing routines, expectations etc. that naturally exist in the place of any real world ubicomp deployment.

## 1 Introduction

In Weiser's seminal paper [20] introducing ubiquitous computing he described how:

*"The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it".*

Weiser then goes on to describe how:

*"...the idea of a "personal" computer itself is misplaced and that the vision of laptop machines, dynabooks and "knowledge navigators" is only a transitional step toward achieving the real potential of information technology. Such machines cannot truly make computing an integral, invisible part of people's lives. We are therefore trying to conceive a new way of thinking about computers, one that takes into account the human world and allows the computers themselves to vanish into the background."*

So, in order for technologies to effectively disappear they need to be carefully designed such that they cause a minimum of disruption with the existing practices and so forth of a given setting (in the human world), e.g. a domestic setting such as a family home or a particular office/work setting.

Talking about settings in this way, it is important to regard them as places and according to Harrison and Dourish [10] a place is:

*“A space which is invested with understandings of behavioural appropriateness, cultural expectations, and so forth”*

Consequently, if the practices afforded or imposed by the ubicomp technology deployment do not fit in with the cultural expectations, existing patterns of behavior etc. associated with a given place then problems with the technologies adoption are more likely to occur.

As a designer/developer of ubicomp technologies it is therefore essential to understand the social and physical richness of a given setting. Typical approaches for understanding settings include ethnographic studies, use of cultural and technology probes, focus groups and design workshops. It would appear that methodologies with a strongly iterative nature, i.e. those based upon a cycle of: observe, design and deploy, observe... etc. are effective for both understanding a given setting prior to deployment and for understanding the adoption of the deployed technology in the setting and the need for design/deployment modifications to occur.

Another factor which is likely to increase the likelihood of a successful ubicomp deployment is to follow a design methodology that places a strong emphasis on end-user involvement – so called, user centered design methodologies. Indeed, through approaches such as participatory design, the end users themselves may be strongly involved in the design process itself.

In the remainder of this chapter, I will focus on our research activities on exploring a particular kind of ubicomp technology – that of situated displays. When we refer to situated displays we agree strongly with the definition provided by O’Hara and his colleagues [15] and the possibilities they raise:

*“At their most basic, digital display technologies allow information to be more easily updated dynamically and remotely. However, these new kinds of interaction technologies also allow people to use these situated displays in novel ways both as for the individual’s purposes and in the support of group work.”*

But it is important to note the emphasis of the word situated which implies a need to take particular notice of the place of deployment when considering this particular form of digital technology.

In exploring situated displays we have three broad research aims:

1. *Understanding of Settings.* In our research we use ethnographic and related studies (both longitudinal and short term) to understand the social nature of public and semi-public spaces both before and after the introduction of situated display technology. This work involved developing an understanding of the affordances of a given place (e.g. outside an office door or inside a communal living area) to help determine appropriate placement strategies for situated displays and an appreciation of what content may be relevant to display at a given place to facilitate cooperation (and sense of community) between and within a certain user group.
2. *Exploration of Interaction and Use.* Situated displays do not typically fit the traditional single user mouse/keyboard interaction style. Consequently, the

project sought to explore the interactions that manifest themselves (over time) in the settings studied. Much of this exploration was guided on our understanding of the settings and utilised techniques found in context-aware computing (location-aware behaviour, automatic personalisation/content creation based on sensed context, etc.) and tangible interfaces as well as more familiar modalities such as e-mail, instant messaging and mobile phones.

3. *Prolonged Deployment.* A key element of the project's research methodology was the use of substantial deployed installations. The long term use of novel technologies, especially their collaborative and community effects, cannot be deeply understood through short-term experiments or 'toy' installations. This development and deployment enables longitudinal studies as well as being a technology demonstrator for dissemination and inspiration. It is also important to note that all the deployments described in this chapter run 24 hours a day, seven days a week.

Settings that we have studied as part of our exploration of situated displays have included: the homes of lecturing/research staff, a Computing Department space at Lancaster University, a residential care facility, a University climbing club and a rural village near Lancaster.

In the remainder of this chapter, we focus on the situated display deployments associated with the Computing Department setting (Section 2) and the homes of lecturing/research staff (Section 3). Following these two Sections we discuss salience/calmness, a key theme in research on situated displays, in Section 4. Related Work is presented in Section 5 and finally Section 6 contains concluding remarks.

## 2 Situated Displays in the Computing Department Setting

### 2.1 Overview

Before July 2004, the Computing Department at Lancaster University was based over three floors in one of Lancaster University's older buildings (see Figure 1). This was the setting of the ubicomp style deployment of ten Hermes1 digital door displays [2, 3 ,4, 5, 8] and an early version of the Hermes photo display (see Figure 2). Owners of displays included lecturers, research assistants, PhD students and administrative staff.



**Fig. 2.** View along a corridor showing three of the original Hermes1 door displays.



**Fig. 1.** View along the corridor where the first Hermes photo display was situated

Hermes consisted of small PDA-sized door units (in fact constructed from PDAs) placed beside office doors. The units enabled their owners to set and display messages and for visitors to scribble notes on the unit's touchscreen display. The Hermes1 units were deployed on 10 doors for a period of 27 months, from April 2002 until July 2004 and during this period its 775 notes were left by visitors and over 5278 messages were set by owners. They were very successful in providing an understanding of long-term usage and adoption of semi-public displays and of the ways in which people were willing to share context information about themselves to others in this particular kind of setting.

In June 2004 the Computing department was relocated to a new building known as InfoLab21. This led to the design and deployment of a new version of the Hermes door display system known as Hermes2 and a new design for the Hermes photo display. A view along one of the corridors showing a number of Hermes2 displays can be seen in Figure 3.



**Fig. 3.** View along one of the corridors of the Hermes2 deployment.

## 2.1 The Hermes1 Door Display and Photo Display Deployment

One of our goals for developing and deploying the Hermes door display system within the main computing building at Lancaster University was to explore whether some of the traditional methods for sharing personal information, e.g. sticking a post-it note outside one's office door, could be achieved with a digital equivalent – one that might present different properties and/or perceived affordances and encourage or encompass different patterns of use, e.g. remote interaction.

The first Hermes door display was installed outside one of the offices in the computing department is shown below in Figure 4.



**Fig. 4.** An early Hermes1 display.

### Supported Functionality

One of the challenges with Hermes was to avoid implementing every feature technologically possible – instead we chose to foster and facilitate a process of modest incremental design (a document describing in detail the incremental changes to the functionality is available from the supporting web site). Our intention was for Hermes displays to have the ease of use and dependability associated with an information appliance and to perform a small number of tasks simply and well.

The functionality supported by the system can be considered from two main perspectives, namely: the perspective of the owner of the Hermes display and the perspective of a visitor to the Hermes display. Visitors were able to leave the owner a message by handwriting with a stylus on the Hermes display. Based on requests from owners we designed the system such that messages left by visitors did not remain on the screen but disappeared from the display once entered. An owner could read his or her messages left by visitors on a web-page (see Figure 5).



**Fig. 6.** Viewing Messages Left by Visitors



**Fig. 5.** A typical textual owner message left on a Hermes1 display.

Owners could create messages to appear on their Hermes display by using a web interface. A typical textual message is shown in Figure 6 and illustrates how an owner would often use his display as a means of sharing personal context based on either their current activity (as in this example) or personal context related to location (e.g. if the message had read “Gone to Gym”) or personal context related to time (e.g. if the message had read “Back in 5 minutes”). A full analysis of the ways in which the Hermes1 system was adopted and used by owners in order to share personal context information is presented in [2].

While, the web interface enabled owners to upload a graphical image for display, such as an animated GIF, feedback from owners suggested that we should add a feature to enable the owner to create a freehand message by using an interface on the door display itself. An example of the kind of message that was left on door displays once this feature was implemented is shown below in Figure 7.



**Fig. 7.** A typical ‘scribbled’ owner message.

This process did, however, require the owner to authenticate themselves with the system; this would typically be achieved by the owner entering a username/password via a simple GUI on the Hermes display. Owners found the overhead of authentication high and so following discussions with users we introduced a means for enabling users to set a temporary message, such as “Gone for lunch” by tapping twice on their Hermes display. The first tap brings up a set of buttons each one representing one of a set of pre-configured messages. A second tap is then required to select the message of the owner’s choice. This technique for setting temporary messages proved very popular for the majority of Hermes owners. Indeed analysis of usage logs revealed that the average number of messages set per day increased from three to nine following the introduction of this interaction feature. This highlights an important tradeoff between flexibility/control and effort on behalf of the user.

As described earlier, one of the issues that we wanted to explore with the Hermes system was the implication of supporting remote interaction with Hermes displays. Consequently, we designed the system such that (in addition to using the web interface to remotely create a message) the owner could also use SMS or MMS via his or her mobile phone in order to send a message or picture to their Hermes display. This feature of Hermes was extremely useful for enabling a door display owner to post their current status when she was not co-located with her display, for example if stuck in bad traffic on the way to work or otherwise delayed from being at their office. A full analysis of this aspect of the system is described in [5] and [13] but the following example message that was texted to a display:

*“In big q at post office.. Will be a bit late”.*

This message provides a good illustration of how messages texted to the system were used in order to manage the owner’s presence. Indeed the majority of messages sent using the system were associated with the notion of awareness and more analysis relating to this theme can be found in [3].

One of the later but most popular features to be added to the Hermes1 system was one which enabled owners to e-mail messages to their displays. One owner, in particular who was a secretary in the Department had a regular habit of e-mailing to the department mailing list whenever she was going to be away from the office for more than a few hours. On discussing her requirements, it became evident that adding a feature that would let her simply copy her regular e-mail message to an e-mail address associated with her door display would greatly reduce the effort required by her to keep her presence information updated on her door display. Having implemented the feature, the secretary in question made extensive use of the feature.

One common type of picture that door display owners (in particular lecturers and researchers) posted to their door display was pictures of their latest conference venue and this observation led us to consider whether a larger situated display could prove useful that would support the display of such photos. In particular, we were interested in investigating whether or not by placing such a display in a

corridor we would see community usage develop around the display by people with offices situated close to the display.

Consequently, we developed and deployed the first Hermes Photo Display in one of the corridors on the lower floor of the Computing Department building (see Figure 2 and Figure 8 below).



**Fig. 8.** The first Hermes Photo Display.

The hardware used was based on a wireless Phillips smart display (the DesXcape 150DM) which had the advantage of not requiring a video cable to be connected to a server PC (however, a power cable was still required of course).

The display was in place for a period a several weeks and at the end of this period we spoke informally to those people with offices on the corridor shared with the display. The feedback we received confirmed that the people on the corridor (mostly PhD students but also some lecturers) all felt that the community strengthened the sense of community on the corridor given the patterns of use that had developed around the display – namely, people on the corridor sending pictures to the display of places they had visited or humorous content.

## 2.2 The Hermes2 Door Display and Photo Display Deployment

Following the dismantling of Hermes1 in July 2004 and our move to a new department building in Inforlab21 we saw an opportunity to create an even more ubiquitous display deployment. Indeed, a full deployment of Hermes2 displays across two corridors and 40 offices has recently completed. These new Hermes II units were designed based on extensive user studies and consultation and include cameras, microphones and Bluetooth as well as the use of a larger 7 inch widescreen display. This larger screen was chosen by the majority of door display owners from the original Hermes system during a ‘show case’ study in which a variety of display options (based on high fidelity prototypes) were presented to previous owners.

The larger screen area has meant that door displays have enough screen real estate to enable the screen to be divided into ‘visitor’ and ‘owner’ sections (see Figure 9). In the ‘visitor’ section, the owner can decide which of a set of messaging options are available to the visitor. For the door display shown in Figure 9, the owner has chosen to have the following messaging options available to any visitors: “Record a video message”, “Record an audio message”, “Scribble a message” and “Use on-screen keyboard”.



**Fig. 10.** A Hermes2 display showing both owner message and visitor buttons.



**Fig. 9.** A hermes2 display showing the owner message only, in this case the owner has chosen not to have his door display support the functionality enabling visitors to leave messages.

Alternatively, owners can decide to have the entire screen area reserved for their own messages, as shown in Figure 10 below.

Many of the offices in the Computing Department are multiple occupancy and Hermes2 door displays now provide support for shared offices. In this case the GUI of the display is divided into a number of rows with one row per person. For example the display shown below in Figure 11 is used to support a shared office with two occupants.



**Fig. 11.** A Hermes2 display supporting a shared office with two occupants.

One key innovative feature of the approach is the use of network bootable computers enabling large numbers of displays to be powered on or off using simple to use web based tools. Similarly, software version control across multiple door displays becomes trivial and highly scalable - from the perspective of a systems administrator, the process of updating 4000 displays is no more difficult than that of updating 40 displays. Another novel aspect of the approach was that a middleware was developed in order to enable members of the department to develop their own door display applications by building on or modifying a range of prebuilt components (see [8] for more details). This 'sandboxing' approach has enabled masters projects to utilise the Hermes deployment without risk of corrupting the system and jeopardizing the reliability of its day to day use. It is important to note that aspect of the system is again a case of meeting the requirements of this particular setting. In more detail, if the Hermes2 deployment was not situated in the university's Computing Department, but say its History Department, then developing the system to enable owners to write their own applications would not have been necessary.

The Photo Display was also deployed in the new InfoLab setting. The display itself was mounted using a wooden structure as shown in Figure 12.



**Fig. 12.** A visitor to the Computing Department downloading an image from the Photo Display onto her mobile phone (March 2006).

This version of the photo display supports multi-user interaction for users with Bluetooth equipped mobile phones. In more detail, a user can use their mobile to upload a picture to the display or to select and then download a photo image from the display to his or her mobile phone. Again the Bluetooth send and receive feature, while appropriate for this deployment setting, would not necessarily be appropriate for deployment in a rural setting where typical user's would be much less familiar with the use of Bluetooth (see [13] for more discussion).

### 3 The Family Homes of Lecturing Staff

The Hermes@Home system is a version of the Hermes system that has been tailored for deployment in the home. The layout of the display is similar to that of the Hermes2 displays but instead of buttons being shown for leaving a message the ‘scribble’ message pane is always displayed. The basic idea behind the displays was to support notions of intimacy between family members when one of the family members was away from home on an extended trip abroad, for example, attending a foreign conference. The Hermes@Home unit would act as a display for content, e.g. photos or text messages, of the person away from home while the at home family members would be able to scribble messages on the unit’s touchscreen display in a very lightweight fashion (the displays were designed to be ‘always-on’ and therefore no booting up of the device would be necessary).

A small number of initial ‘formative’ deployments have taken place and an analysis of use has revealed many similar categories of messages to those encountered with the Hermes deployments (see [19] for more details). A typical home deployment is shown below in Figure 13.



**Fig. 13.** A typical Hermes@Home deployment, situated in the home of a lecturer during a 6 week extended visit to Australia.

It is important to note that the placement of this Hermes@Home unit (as was the case with all deployments) was very carefully chosen by the family members. In more detail, its placement was at a location that was frequently passed by the family member as part of her daily patterns or routines. For example, she would pass the display and notice any new messages waiting for her on her way to the kitchen in the morning to make breakfast and throughout the day and finally, once more, on her way to the family bedroom in the evening.

The main pattern of interaction that would occur around the display was one of the family member checking for new pictures or messages from her partner and scribbling ‘touches’ of intimacy. Examples, of these kinds of messages are shown below in Figure 14.



**Fig. 14.** Typical Hermes@Home messages left by the ‘at home’ family member.

It is important to observe how expressive the scribble style of message leaving can be compared to, for example, a text message.

Interviews with the family members who had displays in their homes revealed how even though many of the messages scribbled onto the Hermes@Home display did not appear to contain much information or appear to require much effort to write there was still concern that the person away would read the messages in a timely manner. For example, one family member commented how she would not want to talk to her partner over the phone until he had read all the messages that she had left since the last time they had spoken. Another family member commented how: “I think it would be good to see what [messages] the other person has read or not”.

The interviews with family members also revealed how the displays appeared to become embedded some form of presence of their partner. For example, one family member commented how:

“*For me it was a bit like a window to where Chris is*”,

and another comments how:

“*When I left I said goodbye to it as a link to u if u get me*”.

#### 4 Analysis of the situated of the displays: Salience and Calmness

The design and use of the situated displays raises many issues but in this Section I would like to focus on two interrelated issues, namely: salience and calmness.

In semiotics, salience [22]:

“*Refers to the relative importance or prominence of a piece of a sign*”

In [21] Weiser talks in depth about Calmness and how it relates to technology. One particular passage from his text is relevant to this chapter:

*“But some technology does lead to true calm and comfort. There is no less technology involved in a comfortable pair of shoes, in a fine writing pen, or in delivering the New York Times on a Sunday morning, than in a home PC. Why is one often enraging, the others frequently encalming? We believe the difference is in how they engage our attention. Calm technology engages both the center and the periphery of our attention, and in fact moves back and forth between the two.”*

The balance between salience and calmness is of critical importance to the appropriate design of situated displays of the form discussed in this chapter.

A message such as “Gone for lunch” shown on an owner’s display should not unduly disrupt or grab the attention of a passerby however it should be of sufficient salience to enable a passerby to be aware of the presence of information relating to the door display owner should the passerby wish to interrogate this information by approaching and focusing on the display.

Similarly, with the design of the photo display, it should not be designed in a way presents pictures in an overly dramatic fashion given the placement of the display in a work setting.

For the Hermes@Home displays the family members themselves were free to pick a place in the home where this balance between salience and calmness could be managed. Of course, where the balance lies is an individual matter of preference. For example, some family members would feel comfortable with the display in the bedroom whereas others would not. Another example being how some family members would wish the display to be in a very private place whereas others would be happy for the display to be situated in a more public area of the house, e.g. in the area inside the house next to the main entrance. Consequently, the appropriate placement (and size etc.) of the device is crucial to achieving the correct balance between calmness and salience.

In effect, this is the very essence of Weiser’s point about disappearing technology and illustrates the importance and interdependence of the setting and the placement of the technology deployed therein and the appropriate design of that technology if the ideal of disappearance is to be achieved.

## 5 Related Work

Foundational research into the issues arising from the use of situated displays to support cooperation between work colleagues was carried out at the Xerox media lab in the early 1990s. In particular, researchers at Xerox deployed and evaluated (over many years) the Portholes shared video space in order to study the potential for supporting coordination between work colleagues through peripheral awareness [6] and to explore the control/privacy issues that naturally arise from the deployment of such a system [7].

One notable example of recent research to have focused on the ways in which situated digital displays can be used to support and foster small communities is the Notification Collage (NC) [9] system. This groupware system, developed and evaluated by a small research group at the University of Calgary, enables

distributed and co-located colleagues to post media elements, e.g. sticky notes or video elements, onto a real-time collaborative surface in the form of a large display in a public setting. A similar system to NC is the ‘Screen Saver’ part of the “What’s Happening” system [17]. This system also utilizes a large screen display but to display collages of relevant web pages in an aesthetically sensitive way in order to support a sense of community across a geographically dispersed college of computing at the Georgia Institute of Technology. In [12] the authors describe concepts for and experiences with a Situated Public Display system deployed in a university setting using two kinds of displays, News Displays and Reminder Displays.

As part of their work on the Dynamo multi-user situated display system, Brignull and Rogers [3] discuss the resistance of the public to participate in community activities supported by large displays situated in public areas due, in particular, to social embarrassment. They also consider how groups of people socialize around large public displays and enter into and out of engagement with such displays.

Significant research on the use of situated displays to support community is described in [11]. Of particular relevance to this proposal is McCarthy’s ‘Group Cast’ ‘context-aware’ application that utilises a large public display and presence-sensing technologies in order to display content of mutual interest to work colleagues as they pass the display. In addition, McCarthy developed the ‘OutCast’ system which enables the owner of an office to display content such as personal web pages, public calendar entries, etc. on a medium-sized display touch-screen display situated outside his or her office.

The Hermes systems have a similar role to the OutCast system, enabling owners of door displays to post pictures and messages outside their office doors. Such content may be posted to support coordination with colleagues or may (usually the case with pictures) simply be posted to foster a sense of community in the computing department where the system is deployed. A similar system to Hermes is RoomWizard [15] an interactive room reservation appliance designed to be mounted outside meeting rooms as opposed to offices.

Moving away from the university/research lab domain, a relatively small number of systems have explored the use situated displays to support coordination and facilitate a sense of community for families. In [16] the authors describe the short term trial of a system supporting the sharing of pictures which utilises a laptop-sized display situated in the family’s living room and Mynatt et al [14] developed and evaluated the Digital Family Portrait system in order to explore the potential for supporting remote awareness of infirm family members by sensing their activity through various context sensors and presenting changes to the family member’s overall activity on a picture frame situated in the home of the care giver.

## 6 Conclusion

In this chapter, I have presented my experiences of and perspective on situated displays, a class of system that provides a useful vehicle for exploring the intersection of ubicomp and HCI. In particular, I have attempted to highlight the importance of understanding and designing for the situational context of any ubicomp style deployment in order to increase the likelihood of successfully ‘weaving in’ the technology into its chosen setting.

## References

1. Brignull, H., Rogers, Y.: Enticing People to Interact with Large Public Displays. In: Public Spaces, in Proc. of INTERACT' 03, Zurich (2003)
2. Cheverst, K., Dix, A., Fitton, D., Rouncefield, M.: Out To Lunch: Exploring the Sharing of Personal Context through Office Door Displays. In: Proc. of Intnl Conf. of the Australian Computer-Human Interaction Special Interest Group (OzCHI'03), pp. 74–83 (2003)
3. Cheverst, K., Dix, A., Graham, C., Fitton, D., Rouncefield, M.: Exploring Awareness Related Messaging through Two Situated Display based Systems. In: Special Issue of Human-Computer Interaction 22, 173–220 (2007).
4. Cheverst, K., Fitton D., Dix, A.: Exploring the Evolution of Office Door Displays. In: Public and Situated Displays: Social and Interactional aspects of shared display technologies. K. O'Hara, M. Perry, et al (Eds), pp. 141–169, Kluwer (2003)
5. Cheverst, K., A. Dix, D. Fitton, Friday, A., Rouncefield, M.: Exploring the Utility of Remote Messaging and Situated Office Door Displays. In: Proc. of the fifth ACM International Symposium on Human Computer Interaction with Mobile Devices and Services (MobileHCI '03), Udine, Italy, LNCS, Springer-Verlag (2003).
6. Dourish, P., Bly, S.: Portholes: Supporting Awareness in a Distributed Work Group. In: Proc. of CHI '92, ACM Press (1992).
7. Dourish, P.: Culture and Control in a Media Space. In: Proc. of European Computer-Supported Cooperative Work Conference, ECSCW'93, pp. 125–137 (2003)
8. Fitton, D.: Exploring the Design, Deployment and Use of Hermes: A System of Digital Interactive Office Door Displays. PhD thesis, Lancaster University (2006)
9. Greenberg, S., Rounding, M.: The Notification Collage: Posting Information to Public and Personal Displays. In Proc. of CHI'00, ACM Press (2000)
10. Harrison, S., Dourish, P.: Re-place-ing space: the roles of place and space in collaborative systems. In: Proc. of CSCW '96, ACM Press, pp. 67–76 (1996)
11. McCarthy, J.: Providing a Sense of Community with Ubiquitous Peripheral Displays, in Public and Situated Displays: Social and Interactional aspects of shared display technologies. K. O'Hara, M. Perry, et al (Eds), pp. 283–308, Kluwer (1993)
12. Müller, J., Paczkowski, O. and Krüger, A.: Situated Public News and Reminder Displays. In: Proc of European Conference on Ambient Intelligence, Lecture Notes in Computer Science, pp. 248–265 (2007)
13. Müller, J., Cheverst, K., Fitton, D., Taylor, N., Paczkowski, O., and Krüger, A.: Experiences of supporting local and remote mobile phone interaction in situated public display deployments. In: International Journal of Mobile Human Computer Interaction

- (IJMHCI) special issue on Advances in Evaluating Mobile and Ubiquitous System (to appear)
- 14. Mynatt, E.D., Rowan, J., Craighill, S., Jacobs, A.: Digital family portraits: Providing peace of mind for extended family members, in Proc. of CHI'01, ACM Press (2001)
  - 15. O'Hara, K., Churchill, E., Perry, M., Russell, D., Streitz, N.: Public, community and situated displays: Design, use and interaction around shared information displays. Workshop at CSCW (2002)
  - 16. O'Hara, K., Perry, M. and Lewis S., Situated Web Signs and the Ordering of Social Action. In: Proc. CHI, ACM Press, pp. 65–72 (2003)
  - 17. Zhao Q. A. , Stasko, J.: What's Happening?: Promoting Community Awareness through Opportunistic, Peripheral Interfaces. In: Proc. of Advanced Visual Interfaces , ACM Press, pp. 69–74 (2002)
  - 18. Romero, N., Baren, J., Markopoulos, V., Ruyter, B., IJsselsteijn, W.: Addressing Interpersonal Communication Needs through Ubiquitous Connectivity: Home and Away. In: Proc. of European Symposium on Ambient Intelligence, pp 419–429 (2003)
  - 19. Saslis-Lagoudakis, G., Cheverst, K., Dix, A., Fitton, D., and Rouncefield, M.: Hermes@Home: Supporting Awareness and Intimacy between Distant Family Members. In: Proceedings of the 2006 Australasian Conference on Computer-Human Interaction (2006).
  - 20. Weiser, M.: The Computer for the 21st Century. *Scientific American* 265, 66–75 (1991)
  - 21. Weiser, M., Brown, J.S.: Designing calm technology. In: *Power Grid Journal*, 1, 1 ( 1996).
  - 22. Wikipedia Salience Definition: [http://en.wikipedia.org/wiki/Salience\\_\(semiotics\)](http://en.wikipedia.org/wiki/Salience_(semiotics))

# Evolution of Geometric Figures from the Euclidean to the Digital Era

Partha Bhowmick

Computer Science and Engineering Department  
Indian Institute of Technology  
Kharagpur  
India  
`pb@cse.iitkgp.ernet.in; bhowmick@gmail.com`

**Abstract.** Recent trends from the Euclidean to the digital geometry in solving various problems on the digital plane are presented in this paper. The notional difference of digital geometry with the Euclidean and the allied geometries has also been pointed out to show how the problems are conceivable in the digital paradigm. Significant contributions in solving these problems using number theory, theory of words, and theory of fractions in general, and digital geometry in particular, have been briefed. The paper is mainly focused on digital straightness and digital circularity, with their related problems, theories, and different perspectives in solving various geometric problems in the digital domain, such as analysis, characterization, segmentation, and approximation.

## 1 Introduction

The earliest known systematic discussion of geometry is Euclidean geometry, which is built on five intuitively obvious axioms, and on proving many other propositions/theorems from these axioms. Although many of these results had been worked out by earlier Greek mathematicians, Euclid was the first to arrange these propositions into a comprehensive deductive and logical system. Hence, in fact, Euclid's *Elements* is by far the most celebrated mathematical work of classical antiquity, and stands out as the world's oldest continuously used mathematical textbook. Interestingly, starting with the fundamental propositions of plane geometry, the Elements mostly lays out geometric methods to obtain the fundamentals of number theory, primality and infinitude of prime numbers, arithmetic theory of proportion, geometric series, incommensurable numbers, etc. [16].

Until the present century, the word “geometry” had continued to mean Euclid-postulated geometry for the last two millennia. Euclid's axioms seemed to be so obvious that any theorem proved from them was deemed to be absolutely true. Today, however, the word “Euclidean geometry” is mentioned specifically to distinguish it from other self-consistent non-Euclidean geometries developed since the early 19th century. It is gradually becoming apparent that Euclidean geometry has certain shortcomings in describing the physical space. From Einstein's theory of general relativity, it follows that Euclidean geometry is an approximation to the geometry of physical space subject to the gravitational field.

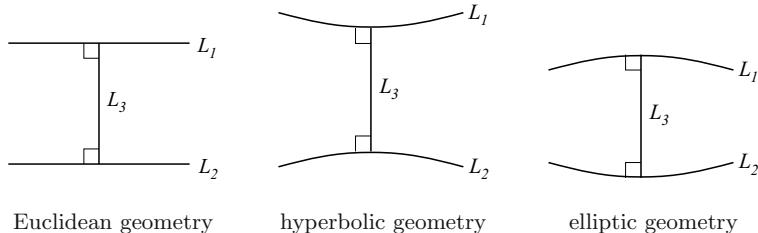
However, Euclidean geometry shows the way to solving problems in different mathematical domains using geometric techniques. It is only the set of axioms and theorems that needs reformulation, which is however the lesser part. The greater part is, whatsoever be the underlying postulates, geometry continues to be a very powerful mathematical science, which, under appropriate convention, would be continuing to serve the noble cause of science and technology with its tremendous potential. In the words of Jules Henri Poincaré, “The axioms of geometry are only definitions in disguise. What then are we to think of the question: Is Euclidean geometry true? It has no meaning. We might as well ask if the metric system is true and if the old weights and measures are false, if Cartesian coordinates are true and polar coordinates are false. One geometry cannot be more true than another; it can only be more convenient.” [39]

## 2 Digital Geometry: A New Discipline in the Digital Era

Although Euclidean geometry is often sought for — quite more than any other non-Euclidean geometry — still today, with the proliferating digitization of graphical objects and visual imageries in the digital era, fresh analytical studies and experimental validation of the simple yet prevalent geometric figures such straight lines, polygons, circles, etc., have become indispensable in order to correlate and harness them for efficient real-world applications. To supplement the interpretation and implementation of these geometric figures in the digital paradigm, digital geometry is seen to have a leading role in recent times [3,4,52,49,50].

Digital geometry deals with discrete sets of points, which are usually obtained from digitized models or images of objects in the 2D or 3D real (Euclidean) space, since digitizing any object means replacing the object by a discrete set of points. Some of its main application areas are as follows:

1. constructing digitized representations of objects, with the emphasis on precision and efficiency [18,19,34,50]
2. analyzing the properties of digital figures, e.g.,
  - (a) digital straightness [35,50,76,77]
  - (b) digital circularity [9,25,48,80]
  - (c) digital planarity [22,21,26,36]
  - (d) digital convexity [41,50,78,95]
3. digital imaging and modeling [56,83]
4. image registration [58,60,68,86]
5. shape analysis [14,71,78,79]
6. biometrics [5,57,61,93]
7. bioinformatics [42,43]
8. medical imaging [58,60,70]
9. document image analysis [11,30,63,94]
10. image and video retrieval [12,13,40,67,87]
11. human motion analysis [37,47,75]
12. vehicle tracking and classification [66,27,92]



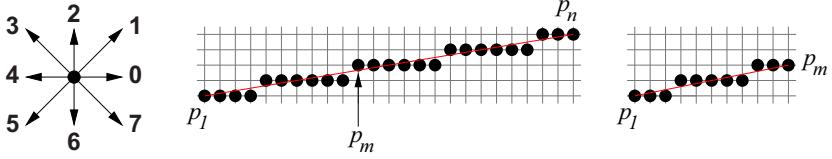
**Fig. 1.** Difference among different geometries apropos the nature of lines with a common perpendicular.

## 2.1 Difference with Other Geometries

Non-Euclidean geometries include hyperbolic and elliptic geometries, which are contrasted with Euclidean geometry by dint of the nature of parallel lines. Euclid's "parallel postulate" (the fifth postulate) states that, for any given line  $L$  and a point  $p \notin L$ , there is exactly one line through  $p$  that does not intersect  $L$ . However, in hyperbolic geometry, infinitely many lines are possible to pass through  $p$  without intersecting  $L$ ; and in elliptic geometry, any line through  $p$  intersects  $L$ . Thus, in short, Euclid's postulates deal with the geometry of flat space, rather than that of the curved space dealt by the non-Euclidean ones.

An illustration on the difference of these postulates is given in Fig. 1. Two straight lines  $L_1$  and  $L_2$  are both perpendicular to a line  $L_3$ . In Euclidean geometry, the lines  $L_1$  and  $L_2$  remain at a constant distance from each other, whatsoever distance one moves further from the points of intersection with their common perpendicular, and are said to be two "parallel lines". In hyperbolic geometry, these lines are called "ultra-parallels", as they "curve away" from each other, increasing in distance as one moves further; and in elliptic geometry the lines "curve toward" each other and eventually intersect.

Digital geometry is different from the aforesaid concepts of both Euclidean and non-Euclidean geometries in its very basis, since a point in the digital plane has integer coordinates only. Such a point is termed as a digital point or an integer point, and also often referred to as a pixel while dealing with digital figures and images. Since a line — whether Euclidean or non-Euclidean or digital — essentially consists of points, a digital line (and any other digital curve) is a sequence of digital points, which satisfies certain straightness properties in a digital sense. Similarities may be found in their constitutions, but differences lie in their very definitions. In the Elements, Euclid defines a (real) point as that of which there is no part. In digital geometry, an additional criterion is that the coordinates of a (digital) point can only be two integers. So, the similarity between a real point and a digital point lies in their very characteristic of having no part, and the dissimilarity is in their coordinate constraints. Similarly, for a real and a digital straight line, the similarity is in the fact that a line is one that lies evenly with points on itself. Interestingly, the property of evenness of points lying on a line was stated as the definition of a real straight line by Euclid in



**Fig. 2.** Left: Chain codes in 8N. Right: DSS( $p_1, p_n$ ) is cut at  $p_m$ ; when the real line segment  $p_1p_m$  is digitized to get DSS( $p_1, p_m$ ), it is not the exact (ordered) subset of points from  $p_1$  to  $p_m$  in the original DSS.

c. 300 B.C., which was reiterated by both Freeman and Rosenfeld [35,50,76] in 1960's-70's in the context of digital straight lines. The difference is, Euclid stated the "evenness" as a definition; whereas, Freeman and Rosenfeld formalized the concept of "evenness" (with a formal proof by Rosenfeld) and had shown that "evenness" is a necessary condition in order that a digital curve is digitally straight. Section 3 explains this in detail. For a digital circle, the similarity with the real circle is its overall "circularity" in the distribution of points. The dissimilarity is, all points of a digital circle are not equidistant from the center of the circle as in the case of a real circle. In fact, opposed to the infinitely many axes of symmetry of a real circle, a digital circle has exactly four axes of symmetry (diameter lines with slopes in  $\{\tan \theta : \theta = k\pi/4 \wedge k \in \{0, 1, \dots, 7\}\}$ ), which creates an asymmetric relation between two digital points in the same octant of a digital circle. This very difference gives rise to challenging problems involving digital circles and digital circular arcs, some of which are discussed in Sec. 4.

### 3 Digital Straightness

Analyzing and interpreting the nature of lines and curves in the digital plane from different perspectives have been an active subject of research since 1960's [50,51,77]. To mention in particular, digital straightness has drawn special attention amongst researchers for their engrossing and challenging nature and for their potential applications related to digital images. For, in a digital image containing one or more objects with fairly straight edges, the set of (exact or approximate) digital straight line segments carries a strong geometric information on the underlying objects, which plays a significant role in an effective abstraction of the underlying objects and in finding the shape-wise similarity between two or more digital objects.

The similarity and the dissimilarity in understanding the very basic nature of a real straight line and a digital straight line have been mentioned in Sec. 2.1. Another important and not-so-trivial characteristic that finds some similarity and also some difference between a real straight line and a digital straight line is as follows. One would expect that, if a digital straight line segment (DSS) is cut into two parts, then each (i.e., sub-DSS) of them would still be digitally

straight (and a DSS, thereof). This, in fact, is true for a real straight line segment in the Euclidean geometry, and also in accordance with our notional intuition. However, from the two subsequences of digital points representing the cut off parts, the correspondence is not straightforward. If  $p_1p_n$  is a real line segment joining  $p_1, p_n \in \mathbb{Z}^2$ , then in the  $\text{DSS}(p_1, p_n)$ , which is the digitization of  $p_1p_n$ , the DSS properties are valid. But if we cut  $\text{DSS}(p_1, p_n)$  at “any” intermediate point, say  $p_m$ , then the (ordered) set of points from  $p_1$  to  $p_m$  (and from  $p_{m+1}$  to  $p_n$ ) might not be the digitization of the real line segment  $p_1p_m$  (and  $p_{m+1}p_n$ ).

For example, in Fig. 2,  $\text{DSS}(p_1, p_n)$  is cut at  $p_m$  in such a way that the right-most run-length of  $\text{DSS}(p_1, p_m)$  is too small compared to the other runs (including the leftmost run-length). In such a case, the new DSS, namely  $\text{DSS}(p_1, p_m)$ , obtained by digitization of the real line segment  $p_1p_m$ , is not a subset of the original DSS, i.e.,  $\text{DSS}(p_1, p_n)$ . Thus, the difference is: When two points  $p$  and  $q$  are arbitrarily selected from a real straight line segment  $L$ , then the new straight line segment  $pq$  is always a part of the original one; whereas, if  $p$  and  $q$  are two arbitrarily selected points from a digital straight line segment,  $\text{DSS}(p_1, p_n)$ , then  $\text{DSS}(p, q)$  may or may not be a part/subsequence of the original segment,  $\text{DSS}(p_1, p_n)$ .

Nevertheless, when a DSS is cut at an arbitrary point  $p_m$  to get a subset  $\langle p_1, p_2, \dots, p_m \rangle$ , it may not be the digitization of  $p_1p_m$ , but it remains to be a DSS in the sense that it represents a part of the digitization of some real line or line segment different from the real line segment  $p_1p_m$ . In fact, a finite set of digital points satisfying the properties of digital straightness is a subset of digitization of infinitely many real lines, which does not stop an arbitrarily cut DSS to remain a DSS. Thus, the similarity is: Whether in case of a real or a digital straight line segment, an arbitrary part of it is always straight — taken in the respective real or digital sense.

The problems related with DSS may be categorized into two classes. One class deals with the mapping from  $\mathbb{R}^2$  to  $\mathbb{Z}^2$ : Given a real straight line segment joining two digital points  $p$  and  $q$ , what is the sequence of digital points constituting  $\text{DSS}(p, q)$ ? There exists several algorithms to find the points of  $\text{DSS}(p, q)$  [17,20,34,50], and the problem is relatively simpler. If  $p = (i_p, j_p)$  and  $q = (i_q, j_q)$ , and w.l.o.g., if  $i_p < j_p$  and if the slope of the real line  $pq$  be in  $[0, 1]$ , then  $\text{DSS}(p, q)$  consists of the nearest digital points corresponding to those real points on  $pq$  which have integer abscissae in the interval  $[i_p, i_q]$ . That is,

$$\text{DSS}(p, q) = \left\{ (i, j) \in \mathbb{Z}^2 : i_p \leq i \leq j_p \wedge j = \left\lfloor y + \frac{1}{2} \right\rfloor \wedge \frac{y - j_q}{j_q - j_p} = \frac{i - i_q}{i_q - i_p} \right\}. \quad (1)$$

The other class deals with problems related with digital straightness, which mainly involves the mapping from  $\mathbb{Z}^2$  to  $\mathbb{R}^2$ . To introduce, some of the intriguing problems from this class are listed below.

1. How many different DSLs (digital straight lines) exist passing through two given points,  $p$  and  $q$ ?
2. How it can be proved that the points of DSS generated by a DSS algorithm are “placed near to” Euclidean straight line  $y = ax + b$ ?

3. How a criterion for initialization conditions can be obtained when DSS extraction algorithm should generate exactly the Bresenham's DSS [17]?
4. Given the sequence of digital points constituting a digital curve, how to decide whether it is a DSS or not?
5. Is it possible to construct any algorithm that can normalize the set (space) of DSS(s), i.e., attaching the length to any DSS starting at a digital point  $p$  and ending at another digital point  $q$ ?
6. How can we extract DSS(s) of maximum possible length(s) from a given digital curve?
7. How can we extract the minimal DSS cover from a given digital curve?

It may be noted that, solutions to Problem 1 and Problem 2 inherit from the close relationship between continued fractions and DSS [53,65,84]. Problem 3 has been addressed in [74], where a DSS has been defined without using the equation of the Euclidean straight line ( $y = ax + b$ ). Regarding Problem 4, in which the input is a digital curve and output is “yes” or “no”, depending on whether or not the digital curve is a DSS, there also exist several interesting solutions [28,31,32,54,64,81]. As evident from their definitions, Problems 5–7 can be classified as very complicated ones. Further, besides the above problems, it is possible to formulate more problems related to DSS [74].

### 3.1 Properties of Digital Straightness

A (irreducible) digital curve  $C$  is a sequence of digital points in 8N or 4N connectivity [50]. In 8N (our consideration),  $(i, j) \in C$  and  $(i', j') \in C$  are neighbors of each other, provided  $\max(|i - i'|, |j - j'|) = 1$ , and the chain codes constituting  $C$  are from  $\{0, 1, 2, \dots, 7\}$  (Fig. 3). If each point in  $C$  has exactly two neighbors in  $C$ , then  $C$  is a *closed curve*; otherwise,  $C$  is an *open curve* having two points with one neighbor each, and the remaining points with two neighbors each. A self-intersecting curve  $C$  can be split into a set of open/closed curves.

In order to determine whether (a part of)  $C$  is straight or not, there have been several studies since 1960’s [50]. Interestingly, solutions to this problem inherit from the theory of continued fractions [50,84]. In [76], it has been shown that  $C$  is the digitization of a straight line segment if and only if it has the *chord property*. (A curve  $C$  has the chord property if, for each point-pair  $(p, q) \in C \times C$ ,  $p \neq q$ , for any  $(x, y)$  on the chord  $\overline{pq}$  (real straight line segment joining  $p$  and  $q$ ),  $\exists(i, j) \in C$  such that  $\max\{|i - x|, |j - y|\} < 1$ .)

From the theory of words [59], it has been shown that, if a DSL is the digitization of a real straight line with rational slope, then the chain code is periodic; otherwise, aperiodic [23]. Formulation of necessary conditions based on the properties of self-similarity w.r.t. chain codes, was first given in [35] as follows:

- F1: at most two types of elements (chain codes) can be present, and these can differ only by unity, modulo eight;*  
*F2: one of the two element values always occurs singly;*

*F3: successive occurrences of the element occurring singly are as uniformly spaced as possible.*

The above three properties were illustrated by examples and based on heuristic insights. Further, Property F3 is not precise enough for a formal proof [72]. Nevertheless, it explicates how the composition of a straight line in the digital space resembles that in the Euclidean space as far as the “evenness” (Sec. 1) in distribution of its constituting points is concerned. A few examples are given below to clarify the idea.

1. 001012100 is not a DSS, since F1 fails (three elements are present).
2. 001001100 is not a DSS, since F2 fails (none of 0 and 1 occurs singly).
3. 010100010 is not a DSS, since F3 fails (singular element 1 is not uniformly spaced).
4. 010010010 is a DSS, since F1–F3 are true.

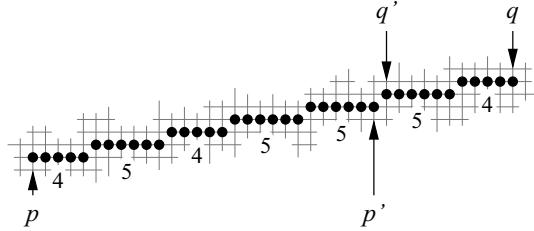
It may be noted that, for a chain code sequence 010100100, the question of “uniform spacing” (as stated in F3) of the singular element 1 remains unanswered. For, there is one 0 between the first two consecutive 1s, and there are two 0s between the next two consecutive 1s. The first formal characterization of DSS, which also brought in a further specification of Property F3, was however provided a few years later in [76], as stated in the following properties of DSS.

- R1: The runs have at most two directions, differing by  $45^0$ , and for one of these directions, the run length must be 1.*
- R2: The runs can have only two lengths, which are consecutive integers.*
- R3: One of the run lengths can occur only once at a time.*
- R4: For the run length that occurs in runs, these runs can themselves have only two lengths, which are consecutive integers; and so on.*

It may be noted that the above four properties, R1–R4, still do not allow a formulation of sufficient conditions for the characterization of a DSS, but they specify F3 by a recursive argument on run lengths. Thus, 010100100 qualifies as a DSS by R1–R4, since the intermediate run-lengths (i.e., 1 and 2) of 0s are consecutive. Another example is given in Fig. 3, which disqualifies Property R4, and hence is not a DSS. Once it is split into two appropriate pieces, the individual pieces pass through R1–R4 tests, and hence each of them becomes a DSS.

### 3.2 Approximate Straightness

As evident from Fig. 3, Properties R1–R4 impose very strict constraints on a digital curve to be a DSS. For real-world digital curves, digital aberrations/imperfections are very likely to occur to certain degree owing to digitization and other preprocessing techniques (e.g., segmentation/edge extraction, thinning, etc. [38]). Hence, straightening of a part of the curve when it is not exactly “digitally straight”, is necessary for practical applications. In a recent work [8], therefore, the concept of approximate straightness has been proposed by relaxing

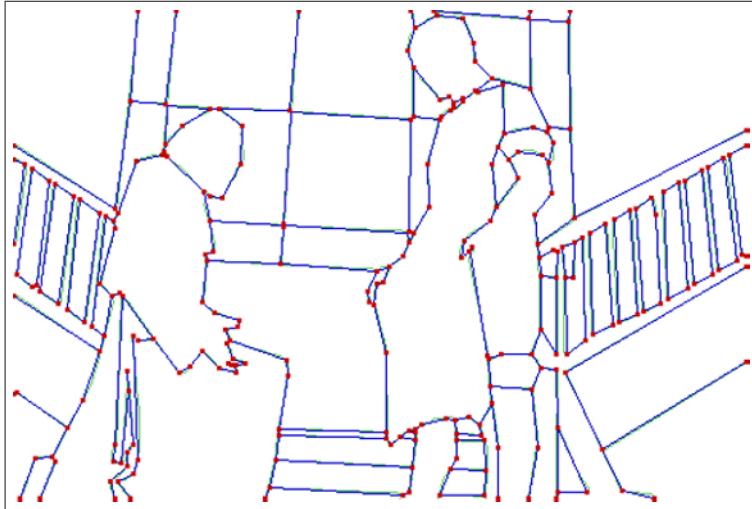


**Fig. 3.**  $C = 0^4 10^5 10^4 10^5 10^5 10^5 10^4$  from  $p$  to  $q$  is not a DSS as it fails to satisfy R4. Its run length code is 4545554, in which the runs of 5 have lengths 1 and 3, which are not consecutive. However, if we split  $C$  into  $C'$  and  $C''$ , respectively from  $p$  to  $p'$  and from  $q'$  to  $q$ , then each of  $C'$  and  $C''$  becomes a DSS.

R2 in a scientific way, and by dropping R3 and R4, so that the output complexity (in terms of the number of extracted straight segments required to cover the curve) becomes significantly less than that of the exact DSS. The extracted set of segments can be used, in turn, to determine a polygonal approximation of the given set of digital curves based on certain approximation criteria and a specified error tolerance,  $\tau$ . Adherence to R1 follows from the fact that it is related with the overall straightness of a digital curve, and a deviation from R1 would destroy the straightness of a line segment.

The notion of approximate digital straight lines segments (ADSS) is based on a set of parameters, which play decisive roles in determining the approximate straightness of a digital curve. One set is the *orientations parameters* consisting of  $n$  (non-singular element),  $s$  (singular element),  $l$  (length of leftmost run of  $n$ ), and  $r$  (length of rightmost run of  $n$ ). For example, the curve in Fig. 3 has  $n = 0$ ,  $s = 1$ ,  $l = 4$ , and  $r = 5$ . Another set is the *run length interval parameters* given by  $p$  and  $q$ , where  $[p, q]$  is the range of possible lengths (excepting  $l$  and  $r$ ) of  $n$ . Greater the relative difference  $(p - q)/p$ , higher is the relaxation achieved for an ADSS.

A sequence  $S$  of ADSS extracted from a digital curve can be used to derive its polygonal approximation based on the adopted merging criterion. Since the set  $S$  provides an elegant and compact representation of digital curves, it is very effective in producing approximate polygons (or, polychains) using the user-specified value of  $\tau$ . The entire algorithm including ADSS extraction and polygonal approximation can be easily implemented, and works efficiently on digital curves having arbitrary shapes and complexities. The algorithm runs very fast compared to other algorithms since it avoids recursion (R4 of DSS) and uses only primitive integer operations. Another advantage is the substantial reduction in the input (sequence of vertices in  $S$ ) size w.r.t. curve length during polygonal approximation. The quality of approximation is ensured by the fact that the sub-optimal approximation does not overstep the worst-case approximation for a given value of  $\tau$ . An approximation on a typical real-world curve set is shown in Fig. 4.



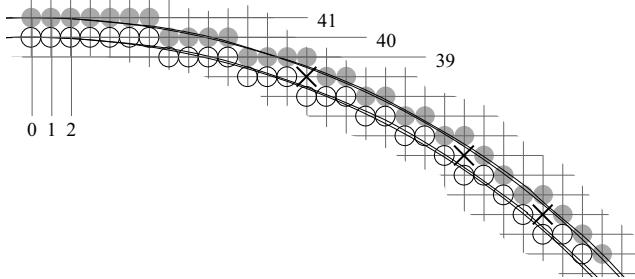
**Fig. 4.** Polygonal approximation of a real-world set of digital curves for  $\tau = 2$ .

## 4 Digital Circularity

Digital circularity, like digital straightness, is also found to possess interesting properties that inherit from digital calculus, number theory, etc. These properties are not only useful to generate digital circles, but also to characterize and recognize digital circles and circular arcs from a given set of digital curves. In order to achieve efficient approximation from the real to the digital domain, digital circles attracted the research community since the adoption of scan-conversion technique [19,24,29,33,44,55,73]. In later years, several works have come up on designing improved algorithms for generation of digitally circular arcs [9,15,20,45,62,82,89,90,91]. However, today's problems related with digital circularity deal mainly with the properties, parameterization, characterization, and recognition of digital circles and circular arcs. A brief discussion of these classes of problems is given here to show the different possibilities where the solutions may be useful in a diverse range of applications.

There exists several definitions of digital circles in the literature, depending on whether the radius and the center coordinates are real or integer values. If we consider the radius  $r \in \mathbb{Z}$  and the center  $c = O$ , then the corresponding digital circle is given by

$$\mathcal{C}^{\mathbb{Z}}(O, r) = \left\{ (i', j') \in \mathbb{Z}^2 : \begin{array}{l} \{|i'|, |j'|\} = \{i, j\} \\ \wedge \\ 0 \leq i \leq j \leq r \\ \wedge \\ |j - \sqrt{r^2 - i^2}| < \frac{1}{2} \end{array} \right\}. \quad (2)$$



**Fig. 5.** Digital circular arcs for  $r = 40$  and  $41$ .

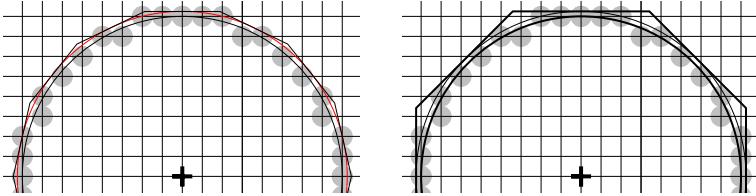
Note that in Eqn. 2, the digital point  $(i, j)$  belongs to the first octant of  $\mathcal{C}^{\mathbb{Z}}(O, r)$ , and hence  $(i', j')$  are the eight symmetric points corresponding to  $(i, j)$ . Rewriting the equation in symmetry-free form, we get

$$\mathcal{C}^{\mathbb{Z}}(O, r) = \left\{ (i, j) \in \mathbb{Z}^2 : \left| \max(|i|, |j|) - \sqrt{r^2 - (\min(|i|, |j|))^2} \right| < \frac{1}{2} \right\}. \quad (3)$$

#### 4.1 Characterization of digital circles

Characterizing the set of continuous circular arcs, which gives rise to a given digitization pattern, has been addressed in [88]. Such a characterization is related with an optimal estimate for the curvature of the continuous arc, which, due to the inherent loss of information in the digitization process, cannot be improved. Hence a method has been suggested in [88] to estimate the local curvature by moving a window along the digital contour. Subsequently, optimal estimation is achieved by considering the domain of all circular arcs that give rise to the specific digital pattern in the window. A domain, which can be one of the six types, namely straight, strictly convex, infinite convex, strictly concave, infinite concave, and non-circular, is calculated for all possible Freeman chains [35] of length  $n$  from 3 to 9. Based on the domain type, the maximal recognizable radius,  $MRR(n)$ , can be estimated to pose a limit on the radius of continuous arcs.

A characterization of the domain of any specific digital pattern in the (center, radius)-space is, therefore, proposed in [88] based on changes in the sequence of curved edges bounding the set of centers in the domain for varying radius. As the radius of elements in the arc domain varies, an unavoidable error is made in the radius or curvature estimation. This error is bounded by the equivalent of the Cramér-Rao bound expressing the influence of digitization rather than stochastic noise. It is called the Geometric Minimum Variance Bound (GMVB), which is computed using the characterization of the arc domain based on a moment generating function. The estimator achieving this minimal variance is the expectation of the shape feature over the arc domain. The method is capable of incorporating prior knowledge on the distribution of the shape parameter of the circular arcs, thereby achieving optimal precision. The shape of the prior



**Fig. 6.** Exact polygonal cover with 14 vertices (left) and approximate polygonal cover with 8 vertices (right) for a digital circle of radius 10.

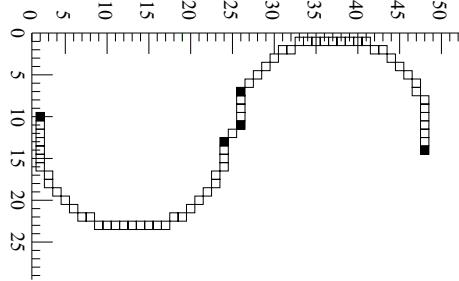
distribution is arbitrary. In the absence of other prior information a uniform distribution of the feature can be assumed. Practical bounds have been given on the precision in measuring curvature or the radius using uniform priors. The relative deviation in the digitization-limited optimal shape parameter measurement of arcs placed at random positions with radius  $r \leq 6$  grid units, using a window of  $n = 9$  Freeman codes, is found to lie between 2% and 9%. For digitizations of full disks (and hence varying  $n$ ) the deviation is below 1% for  $r \geq 4$  grid units.

Results on properties of digital circles in a triangular grid, which are defined by neighborhood sequences based on periodicity, and some new results that do not require periodicity, may be seen in [69]. More recently, in [9], a number-theoretic characterization has been shown based on the relation between perfect squares (square numbers) in discrete (integer) intervals. It is shown how these intervals can be obtained for the given radius  $r$  of a digital circle, and what effects these intervals have on the construction of a digital circle.

Digital circular arcs in the first octant for two consecutive radii,  $r = 40$  and  $r = 41$ , are shown in Fig. 5. The corresponding chain code for  $r = 41$  starting from  $(0, 41)$  is given by  $0^670^370^37070707^307^3$ , in which the runs of 0 usually decrease or continue to be same in length as long as 7 appears as a singular character in the chain code. This is also true for the almost similar chain code corresponding to  $r = 40$ . Interestingly, as shown in [9], the topmost run length for  $r = 41$  is given by the number of perfect squares  $(0, 1, 4, \dots, 36)$  in the interval  $[0, 40]$ , the next run length by that  $(49, 64, 81, 100)$  in  $[41, 120]$ , the next one by that  $(121, 144, 169, 196)$  in  $[121, 198]$ , and so on. Thus, the *square numeric code* for  $r = 41$  becomes  $\langle 7, 4, 4, 2, 2, 2, 1, 1, 2, 1, 1, 1 \rangle$ . By a similar analysis, for  $r = 40$ , we get  $\langle 7, 4, 3, 3, 2, 2, 1, 2, 1, 1, 2, 1 \rangle$ , where the difference occurs when the corresponding intervals contain different number of square numbers, the crossed points in the figure being the points where such a changeover takes place.

## 4.2 Polygonal Approximation of Digital Circles

The problem of covering a (digital) disc/circle by a regular convex polygon in  $\mathbb{R}^2$  has been addressed in [6,10]. For such an *ideal* polygon covering a disc, all the digital points of the disc should lie on and inside the polygon, and vice versa. Approximation of a given digital disc by a suitable regular polygon has significant applications to approximate matching of point sets in the two-dimensional plane



**Fig. 7.** Segmentation of circular arcs into four pieces from an “S”-shaped digital curve.

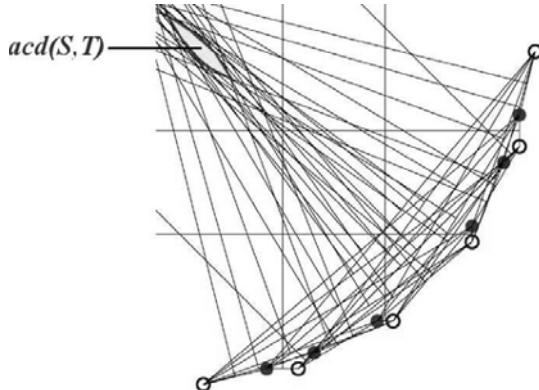
[1,7]. One application is in the domain of biometrics. Given a set  $A$  of digital points, which are the feature points called “minutiae” in a fingerprint image, the task of finding those points (minutiae) in  $A$ , which would belong to a given digital disc, becomes tedious and time-consuming in fingerprint matching [5,46,61,85]. This process of circular range query, however, becomes faster and efficient if we can find a suitable regular polygon in  $\mathbb{R}^2$  (meant for polygonal range query) corresponding to the given digital disc [2,7].

Interestingly, an ideal polygonal cover corresponding to a digital disc is possible for some of the digital discs, especially for the ones having smaller radii. For example, each of the digital circles with radii from 1 to 10 has a regular polygonal cover in  $\mathbb{R}^2$ . The chance of existence of such an ideal polygonal cover, however, goes on decreasing as the radius of the circle gets increasing. For example, there exists no ideal polygonal covers for digital discs of radii 11, 17 – 19, 26 – 31, .... For radius 12 – 16, 20 – 25, 32, 33, 40, the corresponding digital circles possess regular polygonal enclosures.

However, for a digital disc whose ideal regular polygon is not possible (and also for a disc whose ideal polygon is possible), an approximate polygon, tending to the ideal one, is possible, whose error of approximation can be controlled by the number of vertices of the approximate polygon (Fig. 6). Conditions based on which one can decide whether an ideal regular polygon definitely exists corresponding to a digital disc, and conditions under which the existence of an ideal regular polygon becomes uncertain, are explained in [6,10]. Experimental results have also been given to show the tradeoff in terms of error versus number of vertices of the approximate polygon.

### 4.3 Detection/Segmentation of Circular Arcs

Given a digital curve, the problem on segmentation of circular arcs is to partition the curve such that each piece is a part of a digital circle and is locally maximal in length. An instance of segmentation on an “S”-shaped digital curve is shown in Fig. 7. Hough transform [38] forms a standard practice to segment circular arcs, since the method is robust to noisy information in the input data set. However, due its computational burden, which is as high as  $O(n^3)$ , the method is not



**Fig. 8.** The region  $acd(S, T)$  of centers of separating circles [25].

suitable for real-time applications. Hence, improvements have been proposed over the years to detect circular arcs in an efficient way. The two-step algorithm in [48] first computes the centers of the circles based on chord pairs, and then their radius histogram is used for verification and radius estimation. The method avoids any gradient information, as it may be noisy, and uses a threshold value on the 2D accumulator storing the votes of chord pairs.

The work in [25] deals with the digital circle recognition problem in connection with the circular separating algorithm (Fig. 8). It avoids the sophisticated machinery coming from computational geometry or linear programming used in the previous works. It explains how “arc center domain” ( $acd(S, T)$ ) of two given point sets,  $S$  and  $T$ , is computable from the generalized Voronoi cell using the Euclidean metric. An elementary algorithm based on duality (the formal approach to Hough transform) can be used to find a partition of any 8-connected curve in digital circular arcs. The algorithm subsequently helps in estimating the local curvature to a digital curve.

## 5 Concluding Remarks

Some off-late theoretical developments in the domain of digital geometry vis-a-vis the notions of Euclidean and non-Euclidean geometries, related with various domains of computer vision, image processing, and computer graphics are presented in this paper. It is discussed how appropriate modelings and reformulations may be done to conceive the problems in the digital plane, and how to envisage new techniques to solve these problems using interdisciplinary paradigms like digital imaging, number theory, computational geometry in general, and digital geometry in particular. The emerging subject of digital geometry, with its immense potential in solving various geometric problems in the digital domain, has so far found many interesting applications to computer vision and image analysis, such as polygonal approximation of digital curves, shape analysis, circular

range query, approximate point set pattern matching, characterization and approximation of digital circles, biometrics, motion analysis, vehicle tracking and classification, etc. We speculate that several other areas, for example, medical imaging and diagnostics, are yet to be looked at to enrich the theory and to explore future application domains in the digital space.

## References

1. P. K. Agarwal and J. Erickson. Geometric range searching and its relatives. In B. Chazelle, J. Goodman, and R. Pollack, editors, *Advances in Discrete and Computational Geometry*, pages 1–56. American Mathematical Society, Providence, 1998.
2. H. Alt and L. J. Guibas. Discrete geometric shapes: Matching, interpolation, and approximation — A survey. Report B 96-11, 1996. Freie Universität, Berlin.
3. T. Asano, R. Klette, and C. Ronse, editors. *Geometry, Morphology, and Computational Imaging*, volume 2616 of *LNCS*. Springer, Berlin, 2003.
4. G. Bertrand, A. Imaia, and R. Klette, editors. *Digital and Image Geometry: Advanced Lectures*, volume 2243 of *LNCS*. Springer, Berlin, 2001.
5. P. Bhowmick and B. B. Bhattacharya. Approximate fingerprint matching using Kd-tree. In *Proc. 17th Intl. Conf. Pattern Recognition (ICPR)*, IEEE CS Press, volume 1, pages 544–547, 2004.
6. P. Bhowmick and B. B. Bhattacharya. Approximation of digital circles by regular polygons. In *Proc. Intl. Conf. Advances in Pattern Recognition (ICAPR)*, volume 3686 of *LNCS*, pages 257–267. Springer, Berlin, 2005.
7. P. Bhowmick and B. B. Bhattacharya. Approximate matching of digital point sets using a novel angular tree. *IEEE Trans. PAMI* (doi.ieeecomputersociety.org/10.1109/TPAMI.2007.70812), 2007.
8. P. Bhowmick and B. B. Bhattacharya. Fast polygonal approximation of digital curves using relaxed straightness properties. *IEEE Trans. PAMI*, 29(9):1590–1602, 2007.
9. P. Bhowmick and B. B. Bhattacharya. Number theoretic interpretation and construction of a digital circle. *Discrete Applied Mathematics*, 156(12):2381–2399, 2008.
10. P. Bhowmick and B. B. Bhattacharya. Real polygonal covers of digital discs — Some theories and experiments. *Fundamenta Informaticae*, page (accepted), 2008.
11. P. Bhowmick, A. Biswas, and B. B. Bhattacharya. Ranking of optical character prototypes using cellular lengths. In *Proc. Intl. Conf. Computing: Theory and Applications (ICCTA)*, IEEE CS Press, pages 422–426, 2007.
12. A. Biswas, P. Bhowmick, and B. B. Bhattacharya. **CONFIRM: Connectivity Features with Randomized Masks** and their applications to image indexing. In *Proc. Indian Conf. Computer Vision, Graphics and Image Processing (ICVGIP)*, pages 556–562, New Delhi, 2004. Allied Publishers Pvt. Ltd.
13. A. Biswas, P. Bhowmick, and B. B. Bhattacharya. Characterization of isothetic polygons for image indexing and retrieval. In *Proc. Intl. Conf. Computing: Theory and Applications (ICCTA)*, IEEE CS Press, pages 590–594, 2007.
14. A. Biswas, P. Bhowmick, and B. B. Bhattacharya. SCOPE: Shape Complexity of Objects using isothetic Polygonal Envelope. In *Proc. 6th Intl. Conf. Advances in Pattern Recognition (ICAPR)*, pages 356–360. World Scientific, Singapore, 2007.

15. J. F. Blinn. How many ways can you draw a circle? *IEEE Computer Graphics and Applications*, 7(8):39–44, 1987.
16. C. B. Boyer. *A History of Mathematics (2nd Ed.)*. John Wiley & Sons, Inc., 1991.
17. J. E. Bresenham. An incremental algorithm for digital plotting. In *Proc. ACM Natl. Conf.*, 1963.
18. J. E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems Journal*, 4(1):25–30, 1965.
19. J. E. Bresenham. A linear algorithm for incremental digital display of circular arcs. *Communications of the ACM*, 20(2):100–106, 1977.
20. J. E. Bresenham. Run length slice algorithm for incremental lines. In R. A. Earnshaw, editor, *Fundamental Algorithms for Computer Graphics*, volume F17 of *NATO ASI Series*, pages 59–104. Springer-Verlag, New York, 1985.
21. V. Brimkov, D. Coeurjolly, and R. Klette. Digital planarity — A review. *Discrete Applied Mathematics*, 155(4):468–495, 2007.
22. V. E. Brimkov and R. P. Barneva. Plane digitization and related combinatorial problems. *Discrete Appl. Math.*, 147(2-3):169–186, 2005.
23. R. Brons. Linguistic methods for description of a straight line on a grid. *Comput. Graphics Image Process.*, 2:48–62, 1974.
24. W. L. Chung. On circle generation algorithms. *Computer Graphics and Image Processing*, 6:196–198, 1977.
25. D. Coeurjolly, Y. Gérard, J.-P. Reveillès, and L. Tougne. An elementary algorithm for digital arc segmentation. *Discrete Applied Mathematics*, 139:31–50, 2004.
26. D. Coeurjolly, I. Sivignon, F. Dupont, F. Feschet, and J.-M. Chassery. On digital plane preimage structure. *Discrete Appl. Math.*, 151(1-3):78–92, 2005.
27. B. Coifman, D. Beymer, P. McLauchlan, and J. Malik. A real-time computer vision system for vehicle tracking and traffic surveillance. *Transportation Research: Part C*, 6(4):271–288, 1998.
28. E. Creutzburg, A. Hübner, and V. Wedler. On-line recognition of digital straight line segments. In *Proc. 2nd Intl. Conf. AI and Inf. Control Systems of Robots*, pages 42–46, 1982.
29. P. E. Danielsson. Comments on circle generator for display devices. *Computer Graphics and Image Processing*, 7(2):300–301, 1978.
30. A. K. Das and B. Chanda. A fast algorithm for skew detection of document images using morphology. *IJDAR*, 4(2):109–114, 2001.
31. L. S. Davis, A. Rosenfeld, and A. K. Agrawala. On models for line detection. *IEEE Trans. Sys., Man & Cybern.*, 6:127–133, 1976.
32. I. Debled-Rennesson and J. P. Reveilles. A linear algorithm for segmentation of digital curves. *Intl. J. Patt. Rec. Artif. Intell.*, 9:635–662, 1995.
33. M. Doros. Algorithms for generation of discrete circles, rings, and disks. *Computer Graphics and Image Processing*, 10:366–371, 1979.
34. J. D. Foley, A. v. Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics — Principles and Practice*. Addison-Wesley, Reading (Mass.), 1993.
35. H. Freeman. Techniques for the digital computer analysis of chain-encoded arbitrary plane curves. In *Proc. National Electronics Conf.*, volume 17, pages 421–432, 1961.
36. Y. Gerard, I. Debled-Rennesson, and P. Zimmermann. An elementary digital plane recognition algorithm. *Discrete Appl. Math.*, 151(1-3):169–183, 2005.
37. M. Gleicher. Animation from observation: Motion capture and motion editing. *Computer Graphics*, 33(4):51–55, 1999.
38. R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley, California, 1993.

39. M. J. Greenberg. *Euclidean and non-Euclidean geometries: Development and history*. W.H. Freeman, 1993.
40. Y.-H. Gu and T. Tjahjadi. Corner-based feature extraction for object retrieval. In *Proc. Intl. Conf. Image Processing (ICIP)*, IEEE CS Press, pages 119–123, 1999.
41. S. Har-Peled. An output sensitive algorithm for discrete convex hulls. *CGTA*, 10:125–138, 1998.
42. F. Hoffmann, K. Kriegel, and C. Wenk. An applied point pattern matching problem: Comparing 2D patterns of protein spots. *Discrete Applied Mathematics*, 93:75–88, 1999.
43. L. Holm and C. Sander. Mapping the protein universe. *Science*, 273(5275):595–602, August 1996.
44. B. K. P. Horn. Circle generators for display devices. *Computer Graphics and Image Processing*, 5(2):280–288, 1976.
45. S. Y. Hsu, L. R. Chow, and C. H. Liu. A new approach for the generation of circles. *Computer Graphics Forum* 12, 2:105–109, 1993.
46. A. K. Jain, L. Hong, and R. Bolle. On-line fingerprint verification. *IEEE Trans. PAMI*, 19:302–313, 1997.
47. A. Jobbágы, E. Furnée, B. Romhányi, L.Gyöngy, and G. Soós. Resolution and accuracy of passive marker-based motion analysis. *Automatika*, 40:25–29, 1999.
48. H. S. Kim and J. H. Kim. A two-step circle detection algorithm from the intersecting chords. *Pattern Recognition Letters*, 22(6-7):787–798, 2001.
49. R. Klette. Digital geometry – The birth of a new discipline. In L. S. Davis, editor, *Foundations of Image Understanding*, pages 33–71. Kluwer, Boston, Massachusetts, 2001.
50. R. Klette and A. Rosenfeld. *Digital Geometry: Geometric Methods for Digital Picture Analysis*. Morgan Kaufmann Series in Computer Graphics and Geometric Modeling. Morgan Kaufmann, San Francisco, 2004.
51. R. Klette and A. Rosenfeld. Digital straightness: A review. *Discrete Applied Mathematics*, 139(1-3):197–230, 2004.
52. R. Klette, A. Rosenfeld, and F. Sloboda, editors. *Advances in Digital and Computational Geometry*. Springer, Singapore, 1998.
53. J. Koplowitz, M. Lindenbaum, and A. Bruckstein. The number of digital straight lines on an  $n \times n$  grid. *IEEE Trans. Information Theory*, 36:192–197, 1990.
54. V. A. Kovalevsky. New definition and fast recognition of digital straight segments and arcs. In *Proc. 10th Intl. Conf. Pattern Recognition (ICPR)*, IEEE CS Press, pages 31–34, 1990.
55. Z. Kulpa. A note on “circle generator for display devices”. *Computer Graphics and Image Processing*, 9:102–103, January 1979.
56. B. Li and H. Holstein. Using k-d trees for robust 3D point pattern matching. In *Proc. 4th Intl. Conf. 3-D Digital Imaging and Modeling*, pages 95–102, 2003.
57. R. Liang, C. Chen, and J. Bu. Real-time facial features tracker with motion estimation and feedback. In *Proc. IEEE Intl. Conf. Systems, Man and Cybernetics*, pages 3744–3749, 2003.
58. B. Likar and F. Pernus. Automatic extraction of corresponding points for the registration of medical images. *Med. Phys.*, 26(8):1678–1686, 1999.
59. M. Lothaire. *Algebraic Combinatorics on Words*. Cambridge Mathematical Library, 2002.
60. J. Maintz and M. Viergever. A survey of medical image registration. *IEEE Engineering in Medicine and Biology Magazine*, 2(1):1–36, 1998.
61. D. Maltoni, D. Maio, A. K. Jain, and S. Prabhakar. *Handbook of Fingerprint Recognition*. Springer-Verlag, New York, 2003.

62. M. D. McIlroy. Best approximate circles on integer grids. *ACM Trans. Graphics*, 2(4):237–263, 1983.
63. V. Märgner, M. Pechwitz, and H. ElAbed. ICDAR 2005 Arabic handwriting recognition competition. In *Proc. Intl. Conf. Document Analysis and Recognition (ICDAR)*, pages 70–74, 2005.
64. J. A. V. Mieghem, H. I. Avi-Itzhak, and R. D. Melen. Straight line extraction using iterative total least squares methods. *J. Visual Commun. and Image Representation*, 6:59–68, 1995.
65. F. Mignosi. On the number of factors of Sturmian words. *Theoretical Computer Science*, 82(1):71–84, 1991.
66. E. Mohanna and E. Mokhtarian. Robust corner tracking for unconstrained motions. In *IEEE Intl. Conf. Acoustics, Speech, and Signal Processing*, pages 804–807, 2003.
67. F. Mokhtarian and F. Mohanna. Content-based video database retrieval through robust corner tracking. In *Proc. IEEE Workshop on Multimedia Signal Processing*, pages 224–228, 2002.
68. D. Mount, N. Netanyahu, and J. Lemoigne. Improved algorithms for robust point pattern matching and applications to image registration. In *Proc. 14th Annual ACM Symposium on Computational Geometry*, pages 155–164, 1998.
69. B. Nagy. Characterization of digital circles in triangular grid. *Pattern Recognition Letters*, 25(11):1231–1242, 2004.
70. J. Panek and J. Vohradsky. Point pattern matching in the analysis of two-dimensional gel electropherograms. *Electrophoresis*, 20:3483–3491, 1999.
71. I. Pavlidis, R. Singh, and N. P. Papanikolopoulos. An on-line handwritten note recognition method using shape metamorphosis. In *4th Intl. Conf. Document Analysis and Recognition (ICDAR)*, pages 914–918, 1997.
72. T. Pavlidis. *Structural Pattern Recognition*. Springer, New York, 1977.
73. M. L. V. Pitteway. Integer circles, etc. — Some further thoughts. *Computer Graphics and Image Processing*, 3:262–265, 1974.
74. I. Povazan and L. Uher. The structure of digital straight line segments and Euclid's algorithm. In *Proc. Spring Conf. Computer Graphics*, pages 205–209, 1998.
75. J. Richards. The measurement of human motion: A comparison of commercially available systems. *Human Movement Science*, 18(5):589–602, 1999.
76. A. Rosenfeld. Digital straight line segments. *IEEE Trans. Computers*, 23(12):1264–1268, 1974.
77. A. Rosenfeld and R. Klette. Digital straightness. *Electronic Notes in Theoretical Computer Sc.*, 46, 2001. <http://www.elsevier.nl/locate/entcs/volume46.html>.
78. P. L. Rosin. Shape partitioning by convexity. *IEEE Trans. Sys., Man & Cybern.*, 30(2):202–210, 2000.
79. P. L. Rosin. Measuring shape: Ellipticity, rectangularity, and triangularity. *Machine Vision and Applications*, 14:172–184, 2003.
80. P. L. Rosin and G. A. W. West. Detection of circular arcs in images. In *Proc. 4th. Alvey Vision Conf., Manchester*, pages 259–263, 1988.
81. A. W. M. Smeulders and L. Dorst. Decomposition of discrete curves into piecewise segments in linear time. *Contemporary Math.*, 119:169–195, 1991.
82. Y. Suenaga, T. Kamae, and T. Kobayashi. A high speed algorithm for the generation of straight lines and circular arcs. *IEEE Trans. Comput.*, 28:728–736, 1979.
83. G. Tian, D. Gledhill, and D. Taylor. Comprehensive interest points based imaging mosaic. *Pattern Recognition Letters*, 24:1171–1179, 2003.
84. K. Voss. Coding of digital straight lines by continued fractions. *Comput. Artif. Intelligence*, 10:75–80, 1991.

85. J. H. Wegstein. *An Automated Fingerprint Identification System*. US Government Publication, Washington, 1982.
86. J. Williams and M. Bennamoun. Simultaneous registration of multiple corresponding point sets. *Computer Vision and Image Understanding*, 81:117–142, 2001.
87. C. Wolf, J.-M. Jolion, W. Kropatsch, and H. Bischof. Content based image retrieval using interest points and texture features. In *Proc. 15th Intl. Conf. Pattern Recognition (ICPR)*, IEEE CS Press, volume 4, pages 234–237, 2000.
88. M. Worring and A. W. M. Smeulders. Digitized circular arcs: Characterization and parameter estimation. *IEEE Trans. PAMI*, 17(6):587–598, 1995.
89. W. E. Wright. Parallelization of Bresenham’s line and circle algorithms. *IEEE Computer Graphics and Applications*, 10(5):60–67, 1990.
90. X. Wu and J. G. Rokne. Double-step incremental generation of lines and circles. *Computer Vision, Graphics, and Image Processing*, 37(3):331–344, 1987.
91. C. Yao and J. G. Rokne. Hybrid scan-conversion of circles. *IEEE Trans. Visualization and Computer Graphics*, 1(4):311–318, 1995.
92. Q. Zang and R. Klette. Evaluation of an adaptive composite Gaussian model in video surveillance. In *Proc. 10th Intl. Conf. Computer Analysis of Images and Patterns (CAIP)*, pages 165–172, 2003.
93. D. Zhang, G. Lu, W. K. Kong, and M. Wong. Online palmprint authentication system for civil applications. *J. Computers Science and Technology*, 20(1):70–76, 2005.
94. H. Zhu, X. L. Tang, and P. Liu. An MLP-orthogonal Gaussian mixture model hybrid model for Chinese bank check printed numeral recognition. *IJDAR*, 8(1):27–34, 2006.
95. J. Zunic and P. L. Rosin. A new convexity measure for polygons. *IEEE Trans. PAMI*, 26(7):923–934, 2004.

# Web Content Mining Focused on Named Objects

Václav Snášel and Milos Kudelka

VSB - Technical University of Ostrava  
Faculty of Electrical Engineering and Computer Science  
708 33 Ostrava-Poruba, Czech Republic  
vaclav.snasel@vsb.cz, milos.kudelka@inflex.cz

**Abstract.** In our chapter we are working within the field of Web content mining. In relation to the user's description of a Web page, we define a new term: Named object. Named objects are used for a new classification of selected methods dealing with mining information from Web pages. This classification has been made on the basis of a survey of published methods. Our approach is based on the perception of a Web page through an intention. This intention is important both for the users and authors of a Web page. Named object is near to Web design patterns, which became a basis for our own mining method, Pattro. The Pattro method is introduced in this work together with a few experiments.

## 1 Introduction

A Web page is like a family house. Each of its parts has its purpose, determined by a function which it serves. Every part can be named so that all users envision approximately the same thing under that name such as living room, bathroom, lobby, bedroom, kitchen and balcony. In order for the inhabitants to orientate well in the house, certain rules are kept. From the point of view of these rules, all houses are similar. That is why it is usually not a problem for first time visitors to orientate in the house. We can describe the house quite precisely thanks to names. If we add information about a more detailed location such as sizes, colors, furnishings and further details to the description, then the future visitor can get an almost perfect notion of what he will see in the house when he comes in for the first time. We can also take an approach similar to description a building other than a family house (school, supermarket, office etc.) Also in this case the same applies for visitors and it is usually not a problem to orientate (of course it does not always have to be the case, as there are bad Web pages, there are also bad buildings).

Let us look at the problem from the other perspective. If we visit a building with a blindfolded person, then we can submit three basic tasks. The first is to find out what the purpose of the building is. The second is to find out what parts (e.g. rooms) the building contains. And the third task can be linked to the furnishings of individual rooms. When solving these tasks, it is probably possible to start with any of them. There is one more important issue. If the visitor completes some of

the tasks and we will require him to describe the result, he will certainly use commonly used names, which describe the type of building, its parts and finally, its furnishings.

Architect Christopher Alexander [1] brought in a similar and to a certain extent formalized way of description. In our chapter in this book, we try to work with a Web page in a similar way. And we try to show that this way of looking at a Web page can moreover, be a good tool for the classification of some approaches in the field of Web content mining. We can verify it is reasonable to use individually named (labeled) part of a Web page to describe entire page. This holds true both for the suggestion of methods for page semantics detection and for the technically utilizable user's page description.

Our chapter is organized in the following way. In the second section, basic principles concerning Web usability are described. In the third section, we will explain what is meant by Web content mining and what typical tasks are dealt with in this area. In the fourth section of the chapter, we will explain in detail what is meant by the term intention in relation to Web page content. In this section, we will also introduce a new term: Named object, as a basic abstraction related to the intention. The fifth section of the chapter is devoted to the survey of approaches which in some way relate to our view on a Web page. In the sixth section, we present our method Pattrio, which is focused on the detection of Named objects. We will describe experiments related to the successfullness of this method's usability and to its results for partial tasks. The last sections of the chapter are devoted to experiments, a summary and prospects for further research.

## 2 Web Usability

Web usability is closely linked to User Centered Design (UCD). In a wider sense, UCD is a philosophy which results in the process of software system development. The main difference from other approaches is that UCD tries to optimize the user interface, so that it

- Corresponds to what users are used to
- Does not make the user change their way of working

Jacob Nielsen defines usability as follows: "Usability is a quality attribute that assesses how easy user interfaces are to use." In the book [29] many tests with users are shown and many important recommendations for Web page creators come out of the results of these tests.

However, one of the problems is that recommendations in the field of Web usability are not completely formalized. In [14] a conclusion is formulated that recommendations resulting from Web usability can be formalized with the use of patterns (for more about patterns see section Web Design Patterns). One of the pattern characteristics which describe verified experience is that they have an apt name that characterizes the solved task. The methodology of using patterns for Web design is very thoroughly elaborated in [8]. There we can find classification of patterns and important recommendations for developers.

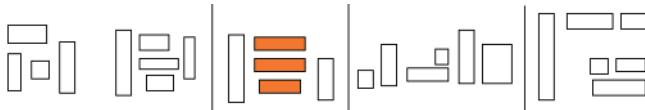
Additional interesting pieces of information related to our view on the Web page are stated in [37]. These pieces of information are related to page layout and parts of page perception. The theory of grouping and alignment was developed early in the 20th century. The Gestalt psychologists described several layout properties. Some of them seem to be important for our visual systems (Fig. 1).

*Proximity* occurs when elements are placed close together. These elements tend to be perceived as a group. If things are close together viewers will associate them with one another. This is basis for strong grouping of content and controls on a UI.

*Similarity* occurs when elements look similar to one another. These elements tend to be perceived as a group. If two things are the same shape, size, color, or orientation, then viewers will also associate them with each other.

*Continuity* occurs when the eye is forced to move through one element and continue to another element. Our eyes want to see continuous lines and curves formed by the alignment of smaller elements.

*Closure* occurs when elements are not completely enclosed in a space. If enough of the information is indicated, these elements tend to be perceived as a group, the missing elements are filled in automatically: We also want to see simple closed forms, like rectangles and blobs of whitespace that are not explicitly drawn for us.



**Fig. 1.** Gestalt principles (proximity, similarity, continuity, closer)

In paper [10], a set of principles based on Gestalt principles is proposed in a pattern language form. At the lowest level, these principles follow Gestalt principles and describe the essential syntax, which establishes the rules for combining words, shapes, and images. These morphological elements have various visible properties such as color, size, thickness, texture, orientation, and transparency.

Also, eye-tracking is an important field in user interface design [13]. It is about how a user's goal influences the way they read and traverse a Web page, which parts of a page users attend to first, how people react to advertising, where they look first for common page elements, how they respond to text, pictures, and so much more. In [11], the following question is answered: "In which way does the visual organization of the Web pages help to lead the visual exploration for an information retrieval?". An explicit goal can be described by respecting two characteristics:

- It must be compatible with the set of the designer's intentions.
- It must be compatible with the set of the user's potentials.

As a result, methods of Web page analysis have to be based on the relations between human perception, cognitive sciences, and biology. The Web page is usually split up into two common structures:

- The physical structure that expresses the organization into geometrical blocks done with homogeneous characteristics. This is the layout of the document.
- The logical structure that expresses the semantic description of the physical organization, that deals with the human interpretation of each block, such as line, section, title, and paragraph.

### 3 Web Content Mining

Web mining is the usage of data mining technology on the Web [16]. Specifically, it is the case of finding and extracting information from sources that relate to the user's interaction with Web pages. In 2002 [15], several challenges had been formulated for developing an intelligent Web: Web page complexity far exceeds the complexity, the Web constitutes highly dynamic information, the Web serves a broad spectrum of user communities, and only a small portion of Web pages contain truly relevant or useful information source of any traditional text document collection. Therefore, following issues should be solved primarily: mining Web search-engine data, analyzing Web link structures, classifying Web documents automatically, mining Web page semantic structures and page contents, mining Web dynamics, building a multilayered, multidimensional Web, etc. Most tasks are still relevant today.

Mainly three areas of Web data mining are concerned [21]: Web content mining, Web structure mining and Web usage mining. Web content mining describes the discovery of useful information from Web contents, data, and documents. The Web content data consist of unstructured data such as free texts, semi-structured data such as HTML documents, and a more structured data such as data in the tables or database generated HTML pages. The goal of Web content mining is to improve finding information or filtering information for the users. Our research focuses on Web content mining whose purpose is to analyze Web pages to find out which useful information is included in the Web page (from the user's point of view). In this area, two sub-areas can be found. The first one is based on information retrieval and its purpose is to find information useful for finding relevant Web pages in large collections (Web searching). The other one is based on information extraction [5] and its purpose is to find structure information that can be, for example, saved in the database and process it accordingly (for example the name and price of products). We are trying to contribute to this area by the fact that according to us, the usable tool for the area of Web data mining are Named objects and Web design patterns, and while doing it we are touching both these sub-areas (see [24]).

## 4 Intent Abstraction

Every single Web page (or group of Web pages) can be perceived from three different points of view. When considering the individual points of view we were inspired by specialists on Web design [8] and on the communication of humans with computers [3]. These points of view represent the views of three different groups of people who take part in the formation of the Web page (Fig. 2).

1. The first group is those whose intention is that the user finds what he expects on the Web page. The intention which the Web page is supposed to fulfil is consequently represented by this group.
2. The second group is users who work with the Web page. This group consequently represents how the Web page should appear outwardly to the user. It is important that this performance satisfies a particular need of the user.
3. The third group is developers responsible for the creation of the Web page. They are therefore consequently responsible for fulfilling the goals of the two preceding groups.

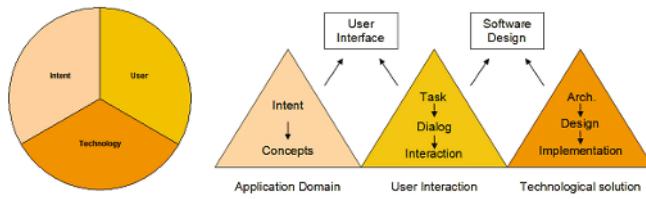
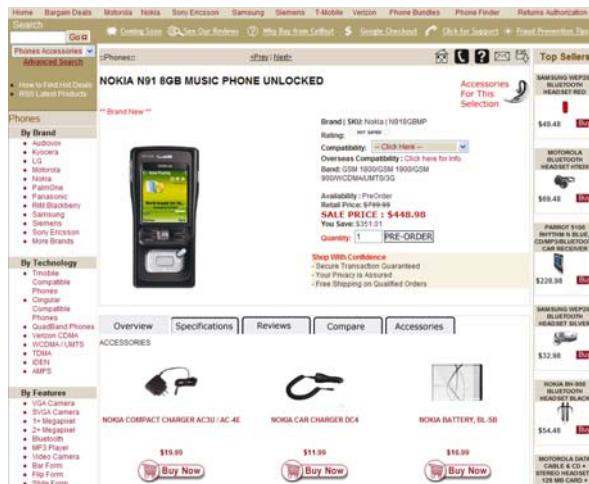


Fig. 2. Web page – user, intention, and technology point of view

### 4.1 User Point of View

If we use an example of a real Web page (Fig. 3), then from the user's point of view it is a typical Web page with a product, where the basic information on the product can be found (including a picture), as well as the possibility of purchase and other offers. The visual aspect and anticipated behavior of the Web page in the domain of product sale can be described as usual, and from this point of view the Web usability is very good. In connection with the point of view of the user who will work with the Web page it can be expected that his needs will be sufficiently satisfied.



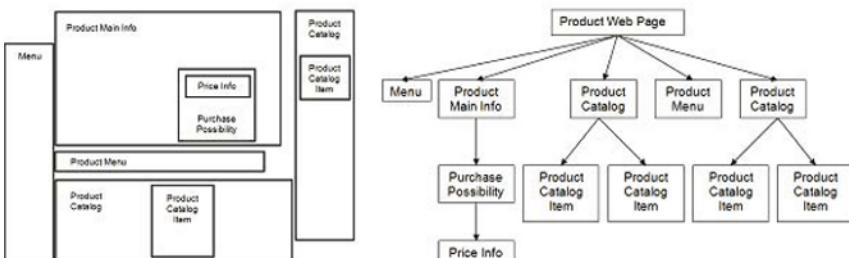
**Fig. 3.** Product Web page

## 4.2 Intention

We can perform a certain generalization in connection with those who define the intention of a Web page. Intention can be described as a series of tasks, which are to be implemented on the Web page. Among those tasks are counted the display of information about the product, the display of a catalogue with constituent items of the accessories, a provision for the possibility to continue to order or purchase the product, a provision of possibilities to navigation to further related Web pages etc. The intention is therefore to provide the user with such a page, where the user gets all information needed for a possible product purchase. In the Fig. 4 the areas representing the aforementioned tasks are simply depicted.



**Fig. 4.** Intentions



**Fig. 5.** Intention hierarchy (a) and (b)

If we go further in the generalization, we can more or less disregard the detailed graphic and content appearance of the Web page. What is left from the intention can be (in correspondence with users' expectations) named and the Web page can be e.g., roughed out at an abstract level (Fig. 5a). In such a way the outlined scheme can become a basis for further pages with offers of different products. Nevertheless, in the schematic description of the intention we can be even more general and we can implement the description e.g. in the form of a tree (Fig. 5b). In this case the scheme is focused more technically and is fully distinguished from the graphic appearance of the Web page. Besides that it creates a hierarchy, which is an important element for the perception of the whole page.

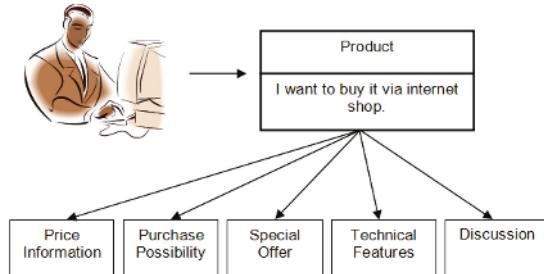
### 4.3 Named object

From the point of view of the user and the developer of the Web page, the linking element is an intention which defines the external architecture and content of the Web page. Whatever the intention is, it is always proven by the appearance of the Web page (overall structure and graphics) and by the type and structure of information contained. To ensure the unified view of users and the author of the Web page in relation to the intention it is important to find a characteristic of the intention which provides a reasonable level of abstraction. This abstraction has to be distinguished from the technical details. As it can be seen in the Fig. 5 (a) and (b), we used a name for every part of the content. This name describes which partial intention is implemented in this part of content. We consider the name to be a key feature of an intention. The concern is actually that both the author of the Web page and the user can relatively unambiguously describe the page with names. The names can serve as a dictionary in this case. In the Fig. 5 (b) is the Web page described more from the point of view of the designer. If we look at the Web page through a task dealt with by the user, we can use following example.

#### Example

The user wants to get to know something about a selected product and possibly to buy it on the Web page. This page can be described in a following way:

“The page contains the *Price information*, the *Purchase possibility* and the *Special offer*. There are also *Technical features* and the *Discussion* at the bottom.”



**Fig. 6.** User's point of view

If we express this description in a graphic form (Fig. 6), then it is similar to the Fig. 5 (b). Apparently the Web page can be described at an abstract level so that the user and the author of the Web page concur. For this description it is necessary to specify more precisely what it is based on. In our considerations it is an abstraction coming out of an intention, which is common to the user and author of the Web page – Intent Abstraction. The level of abstraction can be different when describing the Web page. We can define altogether, three levels.

### Intent Abstraction Levels

1. **Page level.** The first level is based on a type of Web page. This level results from the classification of Web pages, which can be done manually or automatically and it describes the overall intention of the page existence (it therefore does not deal with semantic content). An example of taxonomy can be e.g. Web catalogues [40, 38] or different genre classifications [2].
2. **Domain dependent level.** Second level comes from aforementioned description. There are usually more parts on the Web page and each of them represents a partial target (intention). Individual intentions are dependent on the domain in which the Web page appears. From the point of view of page description, also the semantics of those page elements representing particular targets is very important.
3. **Domain independent level.** The third level is a parallel to the previous level but the intention is not dependent on a domain, but rather it represents, for the user, a different readable characteristic of a page part (e.g. *Menu*, *Advertisement*, *Something to Read*, *Table Data*, *Input Form*, etc.). From the page description point of view, the semantics contained are not important for those elements.

As it follows from the aforementioned, all three selected levels of abstraction have one common characteristic. This characteristic is the name of a descriptive element (or type) of a page. With regard to the fact that the stated descriptive elements are related to the Web page content or its part, we decided to use term *Named object* for them.

**Definition:** Named object (NO) is a part of a Web page

1. Whose intention is general and it repeats frequently
2. Which can be named intelligibly and more or less unambiguously so that the name is understandable for the Web page user

### 4.4 NO Recognition

It follows from previous description, that in order to be able to speak about NO, this element has to be distinguishable by the user. From what attributes should the user recognize, if and what NO there is in question? We work with up to three levels of view:

1. The first view is purely semantic in the sense of the textual content of a page. It does not always have to be a meaning in a sense of natural language such as sentences or paragraphs with a meaningful content. Logically coherent data blocks can still lack of grammars [44]. E.g. for *Price Information* it can be only a group of words and symbols ('price', 'vat', symbol \$) of a data-type (price, number).
2. The second view is visual in a sense of page perception as a whole. Here individual segments of perception or groups of segments of the page are in question. It is dependent on use of colors, font and auxiliary elements (lines, horizontal and vertical gaps between the segments etc.)

3. The third view is a structural one in a technical sense. It is about the use of special structures, such as tables, bookmarks, navigation trees, etc.

If we replace the user's view on a Web page for a computer view, then it is not easy to reach the conclusion that we find algorithms for NO detection on a Web page and for the detection of their semantic content. However, there are approaches, which successfully apply methods implementing these tasks.

## 5 Survey

In our survey of methods dealing with Web content mining we focused on those methods which have their target or at least the object of experiment as this, what we call the Named object. Individual NOs serve us as a tool for the classification of particular methods (see Table. 1).

Let us come back to the metaphor of a family house and let us use NO as a basic descriptive element of a Web page. The computer (algorithm for Web page analysis) is in the same situation as a blindfolded man. The better the page is technically and semantically designed, the easier the job of the algorithm is. Algorithms, which solve for Web pages a parallel of the three tasks from the introduction, can be divided into three groups and we can describe the principles of their functioning in a very simplified way:

**Table. 1.** Named objects

Named Objects	Page Level	Genres (Roussinov, Meyer zu Eissen, Boese, etc.)
		Homepage, Articles, News bulletin, Glossary, Course Lists, Instructional Materials, Geographical Location, Special Topics, Publications, Product Information, Product Lists, Ads, Order Forms, Ratings Help, Article, Discussion, Shop, Portrayal, Hub, Download, etc.
		Web Design Patterns (Page Type) – <a href="http://www.welie.com">www.welie.com</a>
		Article Page, Blog Page, Case Study, Contact Page, Event Calendar, Forum, Guest Book, Help Page, Homepage, Newsletter, Printer-friendly Page, Product Page, Tutorial.
	Domain Dependent	Product Details (Price Information, Purchase Possibility), Special Offer, Product Catalogue, Product Technical Features, Discussion and Comments, Review, Customer Reviews, News, Author and Publications, Book Info, Job Advertisement, Personal Advertisement, etc.
	Domain Independent	Table, Something to Read (Text Content), Link List, Menu, Advertisement, etc.

1. Algorithms for the detection of page type (NO on page level). Methods can focus on semantic features, e.g. keywords (product pages contain words as 'price', symbol '\$' etc.) and data types, or on technical features, e.g. features related to the Web site, which the Web page belongs to (Home Page has a specific URL address, it can contain a Flash presentation) etc.
2. Algorithms for the detection of page parts (NO on a domain dependent or domain independent level). Methods can focus on specific technical features of the Web page (e.g. items of product catalogues are represented by similar DOM trees, such as the text part containing one or more

paragraphs of text with occasional links to pictures) or on the visual layout of a page, to work with semantics search in the found segments and blocs.

3. Algorithms for the extraction of information content of NO (e.g. structured data). Methods require specific procedures aimed at analysis of information content, extraction of data types (attributes) and labels linked to these attributes.

## **5.1 Genre Detection (Intent Abstraction Level: Page)**

The aim of this group of methods is to assign the Web page to some type. This type is either known in advance or it arises as a result of method application. In connection with NO we will deal mainly with the first category.

Genre is a taxonomy that incorporates the style, form and content of a document which is orthogonal to topic, with fuzzy classification to multiple genres [2]. In the same paper there are described existing classifications. Regarding these classifications there are many approaches on genre identification methods. In paper [32], there are analyzed classification problems in terms of two broad textual phenomena: genre hybridism and individualization. The aim of this paper is to show that web pages need a zero to multi genre classification scheme in addition to the traditional single genre classification. The goal of paper [18] is to analyze home page genres (personal home page, corporate home page or organization home page). In paper [4] authors have proposed a flexible approach for Web page genre categorization. Flexibility means that the approach assigns a document to all predefined genres with different weights. In [17] paper is proposed a low-level representation of style of Web pages based on n-grams. Experiments based on two benchmark genre corpora are presented. In [7] paper, there is described a set of experiments to examine the effect of various attributes of web genre on the automatic identification of the genre of web pages. Four different genres are used in the data set (FAQ, News, E-Shopping and Personal Home Pages).

## **5.2 Table Extraction (Intent Abstraction Level: Domain Independent)**

Tables are an important element for structuring related data. The extraction of tables from Web pages appears to be one of the key tasks for further retrieval of structured data. Differentiation of tables emerges as a common problem of all approaches, which represent page layout and those which contain structured data. Tables belong to the field of Domain Independent NO, however, their extraction can serve well for detection of other Domain Dependent NO. A survey of different approaches is in [41]. Generally, tables often have associated text, including titles, captions, data sources, and footnotes or additional text that elaborate on cells. A paper [26] describes an approach to automatic Web table segmentation and extraction of records. This approach is based on the common structure of many Web sites, which present information as lists or tables. In [9] there is proposed an

approach which requires aspects of table understanding, but it especially relies on extraction ontology. An approach describing transformation of arbitrary tables into explicit semantics is presented in [30]. The Tartar system (Transforming ARbitrary TAbles into fRAMES) is based on a grounded cognitive table model. In paper [12] there is presented a different approach. Authors use a tree representation of a variation of the two dimensional visual box model used by web browsers for domain-independent information extraction from web tables.

### **5.3 Opinion, News, and Discussion Extraction (Intent Abstraction Level: Domain Dependent)**

It is a wide area of methods which aim to summarize opinions of customers on a product or on its specific features. In the individual methods, also language analysis (NLP) is used. Opinions of customers on product Web pages are the main source for analysis, but it can also be discussions on thematic forums or individual reviews in the form of articles (NOs Customer Review, Review, Discussions, Blogs). A survey of recent methods of opinion mining is in [25]. In this paper, three tasks specific to opinion mining are analyzed: development of linguistic resources, sentiment classification, and opinion summarization. Another survey is introduced in [6]. The survey is focused on legal blogs (a.k.a. blawg is any Weblog that focuses on substantive discussions of the law, the legal profession, including law schools, and the process by which judicial decisions are made. There is presented a top-level taxonomy shows a variety of topics for blogging sub-community (General legal blogs, Blogs categorized by legal specialty, Blogs categorized by law or legal event, Blogs categorized by jurisdictional scope, Blogs categorized by author/publisher, Blogs categorized by number of contributors, Miscellaneous blogs categorized by topic, Collections of legal blogs. News extraction methods deal with the extraction of articles from news Web pages. From the method point of view it is a special case of methods from the field of Text content extraction. One of the aims of these methods can be to find duplicities published on different Web sites. An approach presented in [31] is based on the concept of tree-edit distance and allows not only the extraction of relevant text from the pages of a given Web site, but also the fetching of the entire Web site content and the identification of the pages of interest. In paper [42] there is a template-independent news extraction approach that simulates human beings. The approach is based on a stable visual consistency among news pages across websites. Discussion as a Named object can be a good source for Opinion extraction. Paper [35] describes techniques to collect, store, enrich and analyze comments on articles. An extraction method is based on similarity of comments and their recurring features (name, date and time, e-mail, etc.). The similar approach is in [27]. The presented system is based on generated wrappers.

### **5.4 Product Details and Technical Features Extraction (Intent Abstraction Level: Domain Dependent)**

Product details is a specific NO on product Web pages. It is a basic characteristic, which usually contains a picture, product name, price information, etc. The aim of the methods is to extract as much information as possible. This information can be saved into a database and then implement operations in this database. An extraction of Product details info is a typical task in supervised or semi-supervised approaches. The goal of a paper [28] is explore a paradigm to enable web search at the object level, extracting and integrating web information for objects relevant to a specific application domain. A similar approach on extracting author metadata is proposed in [43]. Authors attempt to extract author meta-data from their homepages. A paper [34] presents a symbolic approach to extract domain specific technical features of products from large German corpora. The proposed methods depend on manually added lists of technical measures. A similar approach is described in [39] (experiments on camera and MP3 player features). The approach is a two phase framework for mining and summarizing hot items in multiple auction Web sites.

## **6 Pattrio method**

The starting point of our research on Pattrio method is Web design patterns. Generally, the design patterns describe proven experience of repeated problem solving in the area of software solution design. From this point of view, the design patterns belong to key artifacts securing efficient reuse. While the design patterns have been proven in real projects, their usage increases the solution quality and reduces the time of their implementation. Good examples are also the so called Web design patterns, which are patterns for design related to the Web. Even in this area, the patterns are getting quite common (they are collected and published in the form of printed or Internet catalogues, e.g. see van Duyne, welie web). The design patterns are meant for designers and they are free texts containing the problem description, its solution in specific connections and examples of usage. We are trying to look at these patterns from the opposite side (from the point of view of the Web page user) and that enables us to use selected Web design patterns also in a different way.

### **6.1 Web Design Patterns**

Design patterns and pattern languages came from architecture from the work of Christopher Alexander and his colleagues. From the mid sixties to mid seventies, Alexander and his colleagues defined a new approach to architectural design. The new approach, centered on the concept of pattern languages, is described in a series of books [1].

Alexander's definition of pattern is as follow: "Each pattern describes a problem which occurs over and over again in our environment, and then describes the core of the solution to that problem, in such a way that you can use this solution a million times over, without ever doing it the same way twice."

## Forum

### Problem

Users want to discuss a certain topic or react on a particular piece of content on the site.

### Solution

Create a list of topics and allow users to place comments on the topic.

The screenshot shows a forum interface with a sidebar on the left containing links like 'Home', 'About Esato', 'Contact Us', 'Feedback', 'Log In', 'Log Out', 'Create Account', 'Forgot Password', 'Search', and 'Advanced Search'. The main area has a header 'Forum home' with a link to 'www.esato.com/board'. Below the header is a search bar and a message 'The instant support system for your website!' followed by a 'Get Support' button. The main content area displays a list of forum topics:

Topic	Topics	Posts	Last Post
General info	1	1	2002-01-01
News about the Esato site	1	1	2002-01-01
General discussion about Esato products and solutions	4	4	2002-01-01
General discussion about Esato phones	1	1	2002-01-01

Below the table, there is a section titled 'Use when' with the following text: 'You are designing a Community Site or other site for which you are interested in Community Building. Many people can be tied into a site when there is ample opportunity to interact with the site. Such interaction with a site, or its content in particular can also be found on [News Site](#) or on a [Article Page](#).'. There is also a note: 'How A forum is literally a discussions place. It is based on discussion topics and their'.

**Fig. 7.** Sample of forum pattern ([www.welie.com](http://www.welie.com))

According to [37] patterns are structural and behavioral features that improve the applicability of software architecture, a user interface, a web site or something another in some domain. They make things more usable and easier to understand. Patterns are descriptions of best practices within a given design domain. They capture common and widely accepted solutions, their validity is empirically proved. Patterns are not novel, patterns are captured experiences and each their implementation is a little different.

Good examples are also the so called Web design patterns, which are patterns for design related to the web. A typical example of a Web design pattern can be the Forum pattern (see Fig. 7). This pattern is meant for designers who need to implement this element on an independent web page or as a part of another web page. The pattern describes key solution features without implementation details.

## 6.2 Named Object as Web Design Pattern Projection

Patterns are simple descriptions of repeated problems and their solutions. They are intended for developers and they do not contain technical details. That is why their instances in real solutions are not stereotypes and in many cases they are difficult to distinguish. Besides that they deal with different levels of detail (e.g. [www.welie.com](http://www.welie.com)). It can be a level of a whole page, but also e.g. user login, which is just a small part of a page. Some parts of Web pages are on the contrary so easy, that it is not necessary to describe them in a form of patterns. But they can be interesting from a page description point of view.

In our approach, we were inspired by pattern use for the analysis of Web page content. If we look for patterns' instances on Web pages, we will need detailed technical information. That is why we have created own catalogue, in which we describe those repeated Named objects, which we manage to detect on Web pages

by our method. For description we use a description similar to a pattern description, but its intention is different and it aims at understanding what characteristics are important for detection algorithm design. However, our view has a lot in common with patterns. It is mainly because also for us, in the same way as for a pattern, the most important characteristic is the name of the thing described. Our approach is different on the level of the general view and target. Simply said, we understand Named object used by us as a projection of a Web design pattern. This projection does not always have to be unambiguous, e.g. one pattern can be projected to more NOs.

### 6.3 Pattrio Catalog

Patterns are designed for users (Web designers in this case) who work with them and use them in production. A pattern description is composed from parts and each part describes a specific pattern feature. Authors usually use the pattern structure introduced in [1]. In the description there is a pattern name, problem description, context, solution and examples of use. Usually, these are also consequences of the use of the pattern and related patterns which relate somehow with the pattern being used. For our description of NO we use the similar section-oriented structure.

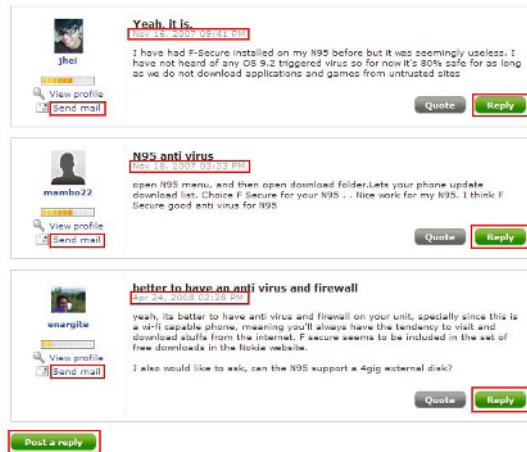
#### Example - Discussion (Forum)

**Problem:** How can a discussion about a certain topic be held? How can a summary of comments and opinions be displayed?

**Context:** Social field, community sites, blogs, etc. Discussions about products and service sales. Review discussions. News story discussion.

**Forces:** A page fragment with a headline and repeating segments containing individual comments. Keywords to labeling discussion on the page (discussion, forum, re, author,...). Keywords to labeling persons (first names, nicknames). Date and time. There may be a form to enter a new comment. Segments with the discussion contributions are similar to the mentioned elements view, in form.

**Solution:** Usually, an implementation using a table layout with an indentation for replies (or similar technology leading to the same-looking result) is used. The Discussion is often together with the Login. If Discussion is on a selling product Website there are usually Purchase possibility, Price information. The Discussion can be alone on the page. In other case there is also the Something to read. In different domains the Discussion can be displayed with Review, News, etc. See Fig. 8.



**Fig. 8.** Discussion

#### 6.4 Detection Algorithm

We have defined sets of elements mentioned above for each NO that are characteristic for this NO (words, data types, technical elements). These elements have been obtained on the basis of deeper analysis of a high volume of web pages. This analysis also included the calculation of weight of individual NO elements that defines the level of relevance for the NO. Besides, we have implemented a set of partial algorithms whose results are the extracted data types and also the score for the quality of fulfillment of individual rules of NOs.

In the context of our approach, there are elements with semantic contents (words or simple phrases and data types) and elements with importance for the structure of the web page where the NO can be found (technical elements). The rules are the way that individual elements take part in the NO display. While defining these rules, we have been inspired by the Gestalt principles. We are using four rules based on these principles. The first one (proximity) defines the acceptable measurable distances of individual elements from each other. The second one (closure) defines the way of creating of independent closed segments containing the elements. One or more segments then create the NO instance on the web page. The third one (similarity) defines that the NO includes more related similar segments. The forth one (continuity) defines that the NO contains more various segments that together create the NO instance.

The basic algorithm for detection of domain dependent NOs [22] then implements the pre-processing of the code of the HTML page (only selected elements are preserved – e.g. block elements as table, div, lines, etc.), segmentation and evaluation of rules and associations (Fig. 9). The result for the page is the score of NOs that are present on the page. The score of the NO then

says what is the relevance of expecting the NO on the page for the user (our experiments show that the relevance of score calculation is approx. 80 % - see [19]).

```

input : Set of PageEntities, set of NamedObjects
output: NamedObjectsScore
1 foreach PageEntity in PageEntities do
2   | if PageEntity is NamedObjectEntity then
3   |   | if does not exist segment then
4   |   |   | create new segment in list of segment ;
5   |   |   | to add page entity to segment ;
6   |   | end
7   |   | add page entity to segment ;
8   | end
9 end
10 foreach segment in list of segments do
11   | compute proximity of segment ;
12   | compute closure of segment ;
13   | compute Score(proximity, closure) of segment ;
14   | if Score is not good enough then
15   |   | remove segment from list of segments ;
16   | end
17 end
18 compute similarity of list of segments ;
19 compute continuity of list of segments ;
20 compute Score(similarity, continuity) of NamedObject ;
21 return Score

```

**Fig. 9.** Detection algorithm

## 7 Experiments

If we are going to work with NO on the user's side, then we can use them as in Fig. 6 as dictionary for page description. If such a description is used, each user can imagine what kind of page it is (without knowing what kind of product it relates to). In the frame of testing of our approach, we have implemented a user interface for a search engine, where we are working with exactly such a page description (see [23]). When a set of results is displayed, each page also contains a tie-on label containing the list of automatically detected NOs. Besides, each page also displays an extended snippet of detected NO instances that is displayed in italics. (see Fig. 10). If our search engine is used, the user can see what type of page it is. The main thing is that pages with the same or similar descriptions are not the same (in the sense of the same design). Furthermore, it is not important if the instances of detected NOs are in the same location. It is really only a description of an external description of page architecture that can say a lot to the user. In our testing interface for search engines, we have implemented the

possibility to re-order the result set according to user requirements. Each page has a tie-on label describing external page architecture by NO. The content of the tie-on label can be understood as a page profile. If the user finds a page whose profile is good for him/her, then the user can click on the tie-on label and reorganize the found set to get the pages with similar profiles into the beginning of this set (our interface is working with the set of the first one hundred pages during reorganization).

**Nokia N95 ProductOverview**  
n95\_productoverview/  
Your source for purchase **Nokia N95**, **Nokia N95** cell phone accessories, **Nokia N95** and **Nokia N95** cell phone accessories best prices.  
... Current price: **\$689.99** ... Availability: In Stock ... Guest Sign-In ... **1900** Mhz and WCDMA  
**2100** Mhz ...  
<http://www.mobilebee.com/Nokia-N95.html>

**Buy Unlocked Nokia n95 Silver Price deals in Canada US UK**  
buy\_unlocked\_n95\_silver/  
Buy **Nokia N95** Silver Unlocked, Price Deals in Canada US UK. Hi-Mobile.Net: Your Online Mobile Phone Shop.  
... USD\$ **689.98** Add to Cart ... **21 mm, 90 cc** Weight **120 g** Display:TFT, 16M colors/240 x 320 pixels, **2.6 inches** ...  
<http://www.hi-mobile.net/Nokia/N95-Fully-Unlocked/>

**Nokia N95 Product Information - PriceRunner UK**  
n95\_product\_information/  
Detailed information about the **Nokia N95** – read the full **Nokia N95** technical specifications.  
... Price range: **E420 - £524.99** ... Internal memory **160.0 MB** The amount of memory ...  
<http://www.pricerunner.co.uk/p/1-741942/Mobile-Phones...>

**Fig. 10.** Tie-on label and snippet

For the purpose of page re-ordering during the interaction of the user with the result set, we need to define how to measure the similarity of web pages. As we have mentioned, the external page architecture is, from the user point of view, understood as a page profile. The user then sees as similar the pages with similar profiles. As we are working with a predefined group of NOs, we can represent the page as a vector of a defined dimension. Currently, we are working with a vector of 24 dimensions; each part of the vector represents the score of one NO. In the case of such a representation we can well use the qualities of the vector-space model (see [33]). The key advantage of this approach is the fact that the dimensions of vectors are not too high. The reason is that the sources of the page description using Web design patterns, genres, and NOs exist as catalogues whose range is limited (to tens of items). This is the reason why the usage of this model is very efficient. The similarity of two pages (and of their profiles) can then be interpreted as the similarity of vectors representing these pages. To compare both vectors, we have used in our experiments the cosine measure representing the cosine of the angle between them.

We have done several experiments whose purpose was to find out if there are any typical web page profiles. For these experiments, we have selected the domain of product sale that is very well covered by existing Web design patterns. We have used various clustering methods [36, 23]. The result of data analysis were three groups of profiles – *Selling*, *Review and discussion*, and *Personal advertising*. We are presenting an example of usage of the SOM [20] – Kohonen map (see Fig. 11)

The clusters found in the data are marked by numbers:

1. Selling: 1, 2, 4, 6, 7, 8, 9, 10, 18, 20, 21, 24, 25 (6,570 pages) - NO: Price information, Purchase possibility, Special Offer, Repayment sale, Technical details.
2. Review and discussion: 3, 16, 17, 22, 23 (3,570 pages) – NO: Discussion, Comments, Review, Technical Details.
3. Personal advertising: 5, 11, 12, 19 (1,050 pages) – NO: Price information, Second Hand, Special offer.
4. Other clusters: 13, 14, 15 (2,310 pages) – pages from other domain (or imperfection of our method).

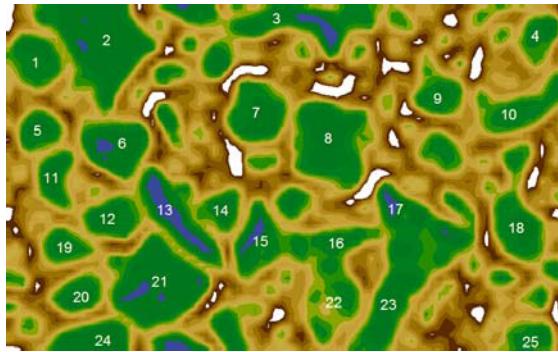


Fig. 11. SOM – product domain clusters

## 8 Conclusion

In our chapter we introduced a new term: *Named object*, which is related to Web page description from the point of view of intention. The intention is common to the authors and the users of a page. NO describes a part of a Web page whose intention can be concisely named. From this point of view we tried to show that two apparently little linked areas have a lot in common. Those are the areas of Web Usability and Web content mining. NO serves as Intent abstraction, with which the page can be described. We performed a survey of methods from the area of Web content mining and we selected those which are related to the use of NO. Selected methods have been classified according to which NO is the subject of use. We also presented our own method Pattrio in the chapter. This method is based on using our own NO catalogue for Web page content analysis, for user Web page description, for searching and for page similarity evaluation. Our experiments have also shown that one of the qualities of NO is their small language and cultural dependence. Our current research leads to the usage of NO for multi-language searching. The replacement of semantically dependent elements in NOs and of algorithms for their extraction can cause our method of NO detection to provide sufficiently relevant results in various language environments.

## References

1. Alexander, Ch.: *A Pattern Language: Towns, Buildings, Construction*, Oxford University Press, New York (1977)
2. Boese, E. S., Howe, A. E.: Effects of web document evolution on genre classification. 14th ACM Information and Knowledge Management (Bremen, Germany, October 31 - November 05, 2005). CIKM '05. ACM, New York, NY, pp. 632–639 (2005)
3. Borchers, J.O.: Interaction design patterns: twelve theses, Position paper, Workshop on Pattern Languages for Interaction Design, CHI 2000 Conference on Human Factors in Computing Systems, pp. 1–6 (2000)
4. Chaker, J., Ounelli, H.: Genre Categorization of Web Pages. ICDM Workshops (2007)
5. Chang, Ch.H., Kayed, M., Girgis, M.R., Shaalan, K.F.: A Survey of Web Information Extraction Systems, IEEE Transactions on Knowledge and Data Engineering, 18, 1411–1428 (2006)
6. Conrad, J.G., Schilder, F.: Opinion mining in legal blogs. Artificial intelligence and Law (Stanford, June 04 - 08, 2007). ICAL '07. ACM, New York, NY, pp. 231–236. (2007)
7. Dong, L., Watters, C.R., Duffy J., Shepherd, M.A.: An Examination of Genre Attributes for Web Page Classification. HICSS (2008)
8. Van Duyne, D.K., Landay, J.A., Hong, J.I. *The Design of Sites: Patterns, Principles, and Processes for Crafting a Customer-Centered Web Experience*. Pearson Education (2002)
9. Embley, D.E., Tao, C., Liddle, S. W. : Automating the extraction of data from HTML tables with unknown structure. Data Knowl. Eng. 5, 3–28
10. Flieder, K., Modritscher, F. Foundations of a pattern language based on Gestalt principles. In CHI '06 Extended Abstracts on Human Factors in Computing Systems, pp. 773–778 (2006)
11. Gagneux, A., Eglin, V., Emptoz, H.: Quality Approach of Web Documents by an Evaluation of Structure Relevance, Proceedings of WDA (2001)
12. Gatterbauer, W., Bohunsky, P., Herzog, M., Krupl, B., Pollak, B.: Towards domain-independent information extraction from web tables. World Wide Web '07 (2007)
13. Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N., Wichansky, A. M.: Eye tracking in web search tasks: design implications. Symposium on Eye Tracking Research & Applications, ETRA '02. ACM, pp.. 51–58 (2002)
14. Graham, L.: *A pattern language for web usability*. Addison-Wesley (2003)
15. Han, J. Chang, K.: Data Mining for Web Intelligence. Computer 35, 11, 64–70 (2002)
16. Han J., Kamber, M.: *Data mining: concepts and techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA. (2000)
17. Kanaris, I., Stamatatos, E.: Webpage Genre Identification Using Variable-Length Character n-Grams Tools with Artificial Intelligence, 2007. ICTAI 2007. pp. 3–10 (2007)
18. Kennedy, A., Shepherd, M.: Automatic identification of home pages on the web. Annual Hawaii International Conference on System Sciences (2005)

19. Kocibova, J., Klos, K., Lehecka, O., Kudelka, M., Snasel, V.: Web Page Analysis: Experiments Based on Discussion and Purchase Web Patterns. IEEE/ACM/WIC Web Intelligence Workshops (2007).
20. Kohonen, T.: Self-Organizing Maps, Springer (2006)
21. Kosala, K. Blockeel, H.: Web Mining Research: A Survey, SIGKDD Explorations, 2, 1–15 (2000)
22. Kudelka, M., Snasel, V., Lehecka, O., El-Qawasmeh, E.: Semantic Analysis of Web Pages Using Web Patterns. IEEE/ACM/WIC Web Intelligence (2006)
23. Kudelka, M., Snasel, V., Lehecka, O., El-Qawasmeh, E., Pokorny, J.: Web Pages Reordering and Clustering Based on Web Patterns. SOFSEM 2008, Novy Smokovec, Slovakia, in Springer LNCS (2008)
24. Kudelka, M., Snasel V., Lehecka, O., El-Qawasmeh, E.: Web Content Mining Using Web Design Patterns, IEEE International Conference on Information Reuse and Integration (2008)
25. Lee, D., Jeong, O., and Lee, S.: Opinion mining of customer feedback data on the web. Conference on Ubiquitous information Management and Communication ICUIMC '08. pp. 230–235 (2008).
26. Lerman, K., Getoor, L., Minton, S., Knoblock, C.: Using the structure of Web sites for automatic segmentation of tables. ACM SIGMOD Management of Data, SIGMOD '04. pp. 119–130 (2004)
27. Limanto, H. Y., Giang, N. N., Trung, V. T., Zhang, J., He, Q., Huy, N. Q.: An information extraction engine for web discussion forums. World Wide Web WWW '05. pp. 978–979 (2005)
28. Nie, Z., Wen, J-R., Ma W-Y.: Object-level Vertical Search. CIDR 2007, pp. 235–246. (2007)
29. Nielsen, J., Loranger, H.: Prioritizing Web Usability. New Riders Press, Berkeley. (2006)
30. Pivk, A., Cimiano, P., Sure, Y., Gams, M., Rajkovic, V., Studer, R.: Transforming arbitrary tables into logical form with TARTAR. Data Knowl. Eng. 60, 567–595 (2007)
31. Reis, D.C., Golher, P.B., Silva, A.S. Laender, A.F.: Automatic web news extraction using tree edit distance. In WWW '04: Proceedings of the 13th international conference on World Wide Web (2004)
32. Santini, M.: Characterizing Genres of Web Pages: Genre Hybridism and Individualization. HICSS 2007, p. 71 (2007)
33. Salton G., Wong, A. Yang, C. S.: A vector space model for automatic indexing, Communications of the ACM 18, 613–620 (1975)
34. Schmidt, S., Mandl, S., Ludwig, B., Stoyan, H.: Product-advisory on the web: An information extraction approach, Artificial Intelligence and Applications, pp. 678–683 (2007)
35. Schuth, A., Marx, M., de Rijke, M.: Extracting the discussion structure in comments on news-articles. ACM international Workshop on Web information and Data Management pp. 97–104 (2007)
36. Snasel, V., Rezankova, H., Husek, D., Kudelka, M., Lehecka, O.: Semantic Analysis of Web Pages Using Cluster Analysis and Nonnegative Matrix Factorization. IEEE/WIC AWIC 2007, Springer ASC (2007)
37. Tidwell, J.: Designing Interfaces: Patterns for Effective Interaction Design, O'Reilly Media, Inc. (2006)

38. Van Welie, M.: Pattern in Interaction Design,<http://www.welie.com>, (last access 2008-08-31).
39. Wong, T-L. W. Lam, W.: Hot Item Mining and Summarization from Multiple Auction Web Sites. ICDM 2005, pp. 797–800 (2005)
40. Yahoo!, <http://www.yahoo.com>, (last access 2008-08-31).
41. Zanibbi, R., Blostein, D., Cordy, J.R: A survey of table recognition: Models, observations, transformations, and inferences, International Journal on Document Analysis and Recognition, 7, 1–16 (2004)
42. Zheng, S., Song, R., Wen, J.-R.: Template-independent news extraction based on visual consistency. In Proceedings of AAAI-2007, pp. 1507–1511 (2005).
43. Zheng, S., Zhou, D., Li, J., Giles, C.L.: Extracting Author Meta-Data from Web Using Visual Features, Data Mining Workshops, ICDM Workshops, 2007, pp. 33–40 (2007)
44. Zhu, J., Zhang, B., Nie, Z., Wen, J.R., Hon, H.W. Webpage understanding: an integrated approach, Conference on Knowledge Discovery in Data, San Jose, California, USA, pp. 903–912 (2007)

# Words and Pictures: An HCI Perspective

Tanveer J. Siddiqui and Uma Shanker Tiwary

Indian Institute of Information Technology Allahabad  
Jhalwa, 211012, Allahabad , U.P., India  
[{tanveer,ust}@iiita.ac.in](mailto:{tanveer,ust}@iiita.ac.in)

**Abstract:** Human and Computer do not speak the same language. This is one of the challenging problems to those working on interface of human and computers. This paper is an effort to summarize the approaches which brings humans and computers a bit closer in terms of interpretation of visual information. When we describe outside world we describe in terms of language expressions but what we see is in terms of pictures. A significant portion of human information is gathered through our visual channel. But we communicate using language (text). However, this is not the case for computational systems. The interpretation of images in computational process is generally in the form of attribute values which cannot be directly correlated with words or concepts. Describing the visual scenes in terms of phrases is the problem in reference for this paper. Recent research efforts focus on combining text and images for semantic image interpretation. We summarize some of these approaches and propose a conceptual framework for information extraction that combines both image and text..

## 1 Introduction

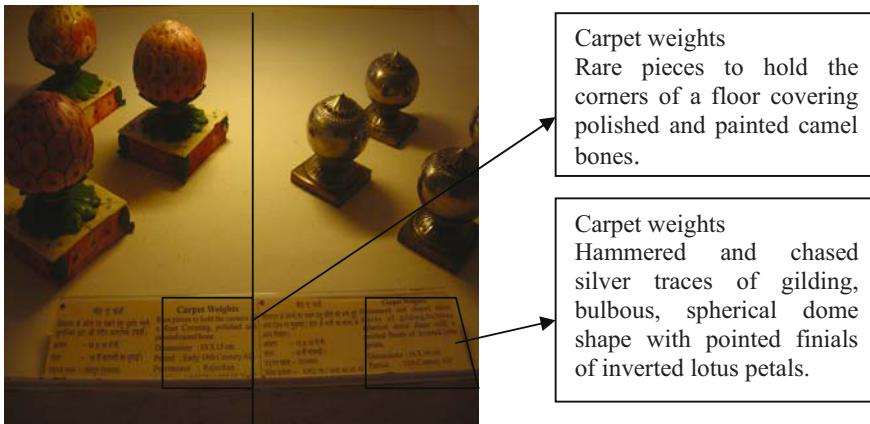
The ability to search and organize information has long been an established research area. The changing nature of the information sources has brought significant changes in it over time. Traditionally information processing has been almost unanimously agreed to mean processing text. Though, the organization, storage and retrieval of information in other media are almost as old as textual information. But when it comes to processing, it was all in terms of text only. Be it a museum dealing with different artefacts or a radio library personnel dealing with recorded tapes. This was achieved with the help of a human annotator who has to annotate every piece of artefact with labels in natural language. It is perhaps the availability of these various modes of information on the web that forced researchers to think about alternative ways of processing. The amount of multi-modal information with which they have to deal now makes it almost impossible to completely rely on human centred approach. The first alternative that came was in terms of processing these other modes of information directly that is dealing with signals (image, speech, etc.) instead of associating them with textual labels. This has lead to the development of image and

speech processing systems. This isolated view of information processing fails to link related information from different modalities together. As a human being we try to correlate various modes of information. Often the retrieval of one mode of information is triggered with another mode. For example, while attending a phone call we can recognize the person. While reading the word ‘lion’ we recall its shape, size, color, food (traits and habits) etc.

The preceding discussion suggests ample reasons for taking an integrated approach for information processing. Research efforts are continued in this direction. Though, a truly brain-like processing is still too far to achieve. Amongst the myriad of articles published in this area we try to summarize the efforts that attempt to combine text and image. Particularly we focus on combining text and images. This integration can extend the capabilities of existing search and retrieval systems, question answering systems and summarization system. Both the text and the image retrieval have long been an established research area. A number of image retrieval systems are already in place. However, efforts to combine them are only of recent interest. The same is true for question answering, summarization systems, etc.

There have been mainly two approaches to image retrieval. The first approach was to annotate images manually with textual descriptions and then using these descriptions in retrieval. Though this approach support semantic queries but suffers from certain disadvantages. First, it is time consuming and expensive. Second, it captures the viewpoint of human annotator (i.e. subjective) and suffers from low term agreement across indexers. Third, it is not always easy to describe the information contained in an image using text because an image not only conveys what is being shown in the image but also what the image is about.

The problem with text-based retrieval has lead to the development of image centred approaches known as content-based image retrieval systems. These systems use visual contents like color, texture, shape etc. in indexing and require user to query based on either visual content or by example. This approach however doesn’t seem natural. People are not familiar to querying images based on visual concepts. They do not get indulge into minute (low level) details of pictures while recognizing it. Instead, they focus on higher level entities and ignore small differences. They would like to query images using textual description. For example, one should be able to pose a query like “Find me images of girls performing ballet” or using a combination of query and textual keywords. In order to handle such queries a CBIR system requires semantic information. This semantic information usually consists of textual descriptions.



**Fig. 1. Carpet weights**

This means that image centred approach can not replace text based retrieval but can complement it. In order to overcome the disadvantages when they are taken separately, we can combine both the approaches. This is achieved by extracting certain features from images while others from accompanying text document. For example, the accompanying text provides useful information about the carpet weights shown in Fig.1. This information can be utilized to enhance the capabilities of CBIR system. Research efforts are going on in this direction. Earlier examples of integrating multimedia data in retrieval framework include [4, 12]. Combining text and images can prove useful in other applications also, e.g. question answering, summarization, etc.

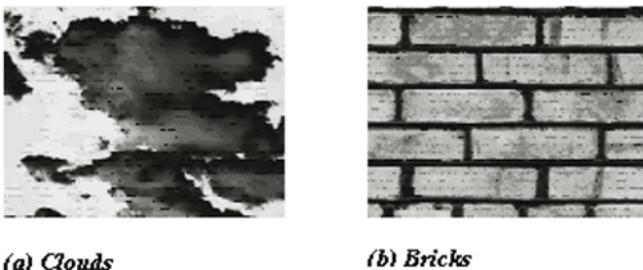
Quite often complete information can not be obtained using image alone or using text alone, electronic health monitoring systems being a prime example where a lot of clinical information remains textual. Images and text can improve image understanding by complementing each other in such cases. This paper summarizes research attempts made in this direction. This combined view of looking at information processing will improve the way of communicating with computers. The rest of this chapter is organized as follows:

The next section introduces content based retrieval systems and summarizes the research efforts to combine text and image in retrieval. Section 3 focuses on how text and images can be combined in a question answering system. Finally in section 4 we propose a generalized framework for combining text and image for extracting information.

## 2 Content-Based Image Retrieval

Content-based image retrieval systems rely on features that can be extracted automatically from images such as color, shape, texture, etc. Retrieving images using color is achieved through the use of color histogram. Spatial relationship among

several color regions has also been used [5]. Texture represents visual patterns present in an image. It can be roughly modelled as a two-dimensional Gray level variation. A two-dimensional dependence texture analysis matrix, called as co-occurrence matrix, is usually calculated to define texture. Texture features include directionality, roughness, contrast, coarseness, likeness, regularity, etc. Interested readers can refer to [11] for further details. Spatial arrangement of objects in an image can also provide useful information in interpreting image semantic. For example, we can not expect a knee bone near heart in a medical image.



**Fig. 2. Examples of Textures**

The focus in CBIR has been on the use of features that can be extracted computationally rather than on visual attributes required by users for various tasks. However, there remains a semantic gap between the digital representation of an image and the interpretation that user associates with it. Image retrieval systems solve this problem using manual annotation of images with keywords. Text annotation means associating keywords to an image based on its visual content. This is a manual process of observing the image and based on its visual content assigning keywords to it.

Google also uses textual annotation of images. The method used in Google involves displaying image to two users, who have to assign as many keywords as possible to it. The common keywords are then identified and attached to the image. Its biggest drawback is that the keywords have to be manually associated and is totally based on user's perception.

With the rapid advancement in imaging technology the amount of image data generated daily has increased significantly. Resorting to manual indexing of images completely doesn't seem to be a practical solution. An alternative to manual annotation is automatic annotation of images. Another direction of improvement in CBIR is combining text and image together for semantic interpretation of images. Knowledge-bases and Ontologies have been used for this purpose [15, 17]. Yet another way of combining content-based and text-based approach to multi-modal retrieval is to perform retrieval separately and then combine the result to improve the performance as in [12]. Their work was focused on a dataset containing medical images and non-structured text. In the following lines we point out some of the earlier work done in the area of automatic annotation and use of knowledge-bases and ontologies for semantic image interpretation.

## 2.1 Automatic image annotation

This problem corresponds to associating words with pictures. One way to achieve this is to use the probability of associating words with image regions as in [2]. Machine learning approaches have been used for this task [2, 3]. [7] has used a vocabulary of blobs to describe images. Each image is a composition of certain number of these blobs. They considered the problem of automatic image annotation as the task of translation from a vocabulary of blobs to a vocabulary of words. The work in [8] described an automatic approach to image annotation and retrieval. A small vocabulary of blobs was used to describe regions in an image. An image I was thus represented as a set of blobs:

$$I = \{ b_1, b_2, b_3, \dots, b_m \}$$

The task of image annotation is to select a set of words  $\{ w_1, w_2, w_3, \dots, w_n \}$  that describes the content of image. They used a probabilistic model to predict the probability of generating a word given the blobs in an image I.

$$P(w/I) \approx P(w/b_1, b_2, b_3, \dots, b_m)$$

The joint probability of observing the word w and the blobs  $b_1, b_2, \dots, b_m$  in the same image was estimated using training set T.

$$P(w/b_1, b_2, b_3, \dots, b_m) = \sum_{x \in T} P(w, b_1, b_2, b_3, \dots, b_m / X)$$

Using the independence assumption this expression was rewritten as:

$$P(w/b_1, b_2, b_3, \dots, b_m) = \sum_{x \in T} P(X) P(w/X) \prod_{i=1}^m P(b_i/X)$$

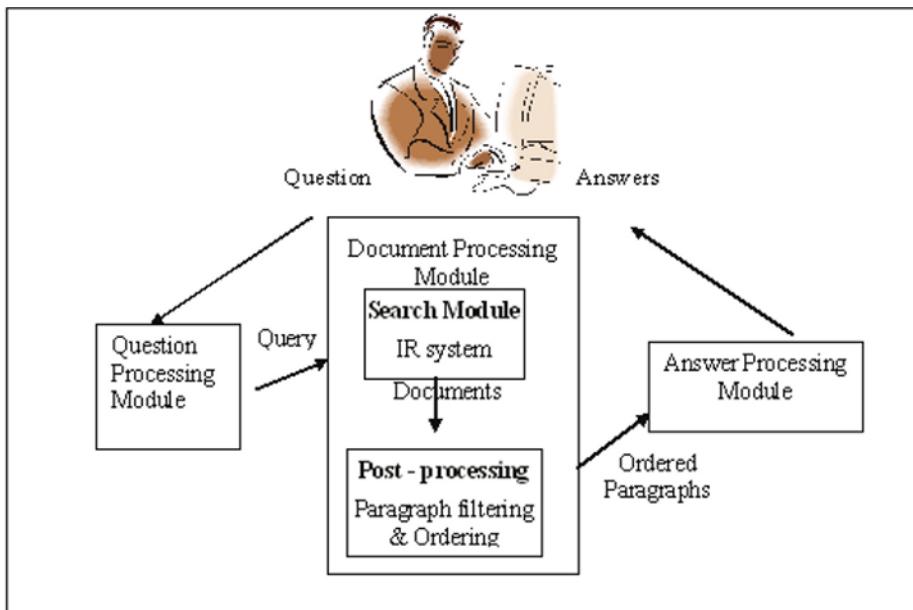
They used smoothed MLE estimates to calculate the probabilities in the above equation.

## 2.2 Using Domain Knowledge

Using domain specific information can help in automatic extraction of meaning from images and thus enabling semantic retrieval. This type of information can be obtained from knowledge bases (KB), ontologies, etc. as in [14,15, 21,[22]. The work in [15] presents a knowledge-based approach to interpret X-ray images of bones. Paper [21] presents a framework that combines syntax and semantic for medical image retrieval. Their framework permits user to query the database using text or image. Keywords in the query are mapped to concepts from ontology. In [22] a machine learning and a knowledge-based approach for mapping between image data and semantic data has been discussed. Town and Sinclair [16] proposed an image retrieval approach based on an extensible ontology. They used supervised ML approach to perform the mapping between the image data and concepts. Visual concept ontology and an image processing ontology have been used in their work.

### 3 Question answering System

Many research and commercial question answering systems have been developed. These systems follow one of the three approaches: Natural language processing (NLP), information retrieval and template matching [1]. NLP-based question answering systems parse the question to understand its syntactic structure then convert it into some semantic representation and use logic to derive an answer. SHRDLU [18], one of the early QA systems, is an example of this type of question answering system. NLP-based QA systems are usually domain-specific in which syntactic analysis of user's questions is intertwined with the semantic interpretation process. An important characteristic of these systems is the integration of domain-specific information. IR-based question answering systems make use of web to answer domain independent questions. These systems use an IR system to find a set of documents in which answer can be found. Answers are then extracted from these documents. Fig. 3 shows the architecture of IR-based question answering system. Ask Jeeves<sup>1</sup> and START (syntactic analysis using reversible transformations) [10] are examples of open domain question-answering systems. Template-based QA system convert questions into a template and then extract facts from documents to fill in these templates.



**Fig. 3. IR-based Question Answering System [23]**

<sup>1</sup> <http://www.askjeeves.com>

Besides these approaches some systems leverage human computation to provide answers, such as Yahoo! answers<sup>2</sup>. These systems maintain a knowledge sharing community where human experts answer the questions on any topic. To answer a question the system searches earlier resolved questions that are similar to it. Mobile companies use this type of QA to answer questions sent by mobile users. Example includes AskMeNow<sup>3</sup> and ChaCha<sup>4</sup>.



**Fig. 4.** Yahoo! Answers interface

All of these systems are based on text alone. These systems find it difficult to handle questions involving objects with distinct visual features. For example,

“Where I can find a show piece like this



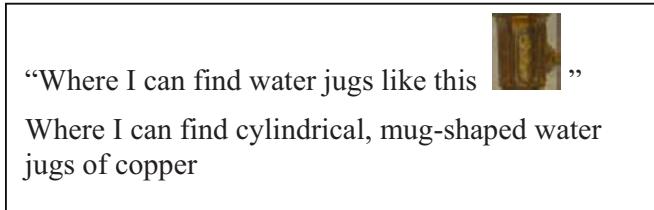
This question is inherently dual-mode consisting of a verbal component and a visual component describing visual attributes. A text-based system requires the user to describe the visual details of the image using text. However, capturing visual properties in text is not an easy task. Even if it is done, it may be hard to find these details in the accompanying text so as to find an answer to the question in hand. In order to answer such questions both the text and image content needs to be utilized. This can be achieved by extending the capabilities of existing question answering systems by providing support for dual-mode questions. This requires integration of existing question answering systems with image matching technologies so as to permit photos to be part of the question. Fig. 5 explain the difference between uni-modal and dual-

<sup>2</sup> <http://answers.yahoo.com>

<sup>3</sup> <http://www.askmenow.com>

<sup>4</sup> <http://www.ChaCha.com>

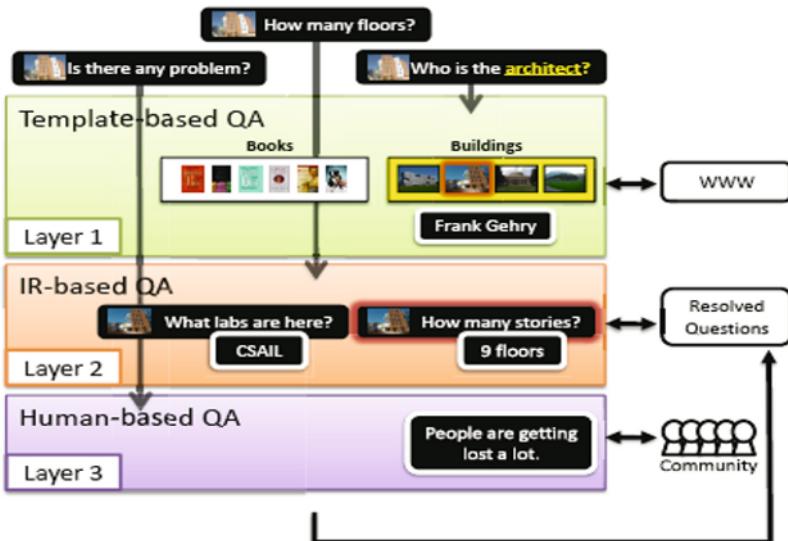
mode questions. In the first case, user provides the link of the poster similar to what she is looking for. Hence, s/he needs not to provide the visual description in the text which the second user has to.



**Fig. 5. Uni-modal and dual-mode questions**

Research attempts involving integration of the two modes in QA include [9, 19]. Yang et al. [19] uses computer vision to answer questions about news video whereas Katz et al. [9] uses activity data captured through real-time motion tracking system to answer activity related questions on surveillance videos. Both [19] and [9] use multimedia data to improve the quality of answers. The question still contains only verbal component (text only) and hence fails to handle question about a particular object, say about a particular car in surveillance videos.

Yeh et al. [20] took these ideas further to include an image component in the question itself. They proposed three-layer architecture for photo-based question answering system as shown in Fig. 6. The three layers in [20] correspond to template-based, IR-based and human-computation based question answering.



**Fig. 6. Three-layer system architecture for photo-based QA proposed in [20]**

The first layer takes a template-based approach to answer questions. The facts to fill in templates are extracted from structured multimedia databases. The second layer takes an IR-based QA to answer questions that first layer fails to answer. It searches a large corpus of already resolved photo-based questions to answer. The third layer seeks the help of human experts to answer much harder questions.

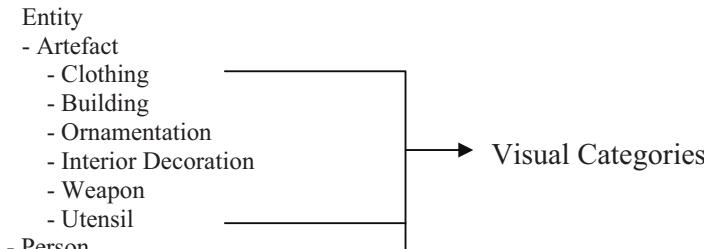
## 4 A Conceptual Framework for Combining Text and Image

Taking cue from cognitive process, a three level information extraction and integration, viz. perception level, attentive (intermediate) level and cognitive level, can be assumed. This three level information extraction model holds for all modalities (text, speech, and image). At the lowest level we try to group low-level details, at the next level we concentrate on some features, and finally information extraction takes place. Human brain effectively utilizes all the modes of information in the cognitive process of building understanding. However, we are not competent enough to discuss at which level fusion across modalities takes place and which modes take precedence over another and when. Some interesting observation has been made in [24].

We here discuss a framework for combining text and images for information retrieval. The image data and the concepts (text) do not share the representation space. This is the major problem in integrating the two modalities. We need to draw the correspondence between these two representations in order to combine them. Given an image, this task can be viewed as a mapping:

$$f: f \rightarrow \{ci\}$$

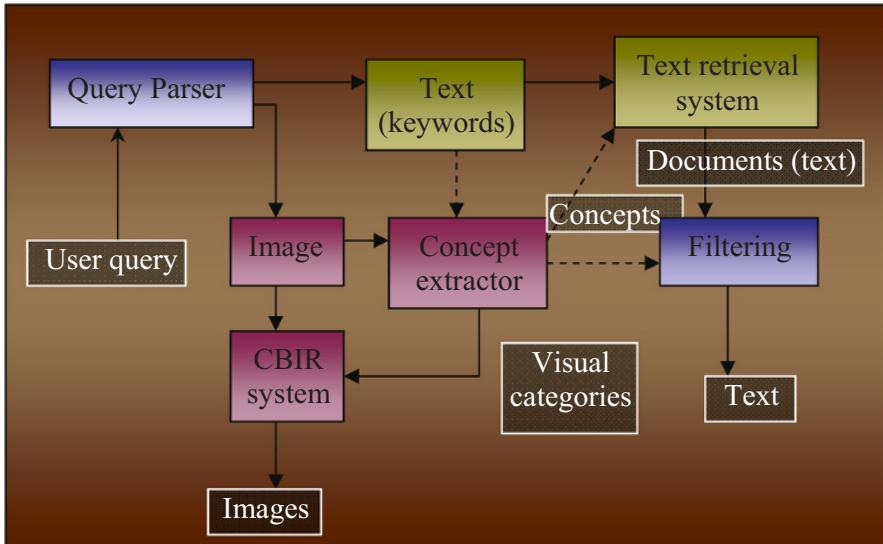
Where  $f$  is image feature vector,  $V_c$  is set of visual categories and  $ci \in V_c$ .



**Fig. 7. A partial image hierarchy**

The information seeking environment considered is of an on-line museum containing images of various artefacts. The application considered is for information extraction/retrieval. These artefacts are accompanied with verbal descriptions (unstructured text). The user seeking information (both text and image) about a particular artefact can query using an image or text or both. We assume the existence of an image hierarchy and a concept hierarchy. A partial image hierarchy is shown in Fig. 7. The visual categories can have sub-categories, e.g. interior decor can have sub-categories like lighting, painting, furniture, etc. A is a type relationship is assumed between a visual category and its sub-categories. Supervised machine learning

approach can be used to create link between low-level image data features and a pre-defined set of visual categories. Training data consists of image of artefacts with manually annotated visual categories and a set of keywords drawn from their textual description.



**Fig. 8 Framework for information extraction**

Fig. 8 shows the framework. A query parser takes the query and separates text and image components. The image component is passed to concept extractor which extracts the visual category of the image and associated concepts. The visual category is used by CBIR to filter retrieved images. If the query consists of image only then these concepts are used to extract concepts. For a dual-mode query we identify top-three visual categories similar to query image and use keywords to filter out irrelevant categories. This filtering is carried out with the help of WordNet concept hierarchy. The hypernym of keywords are matched with visual categories and only matching category is retained. In case of text only query we use hypernym information to identify visual category. A set of images from this category is offered to user to pick up the one similar to intended one. This image is then passed to CBIR system which retrieves images using visual content.

Using the retrieved documents and image we can generate the output in a structured form as shown below:

Entity Type:	Image:	
Description:		
Source:		
Made of:		...

## 5 Conclusion

This chapter summarizes the approaches for combining image and text together for applications like retrieval, question answering, information extraction etc. A conceptual framework has been also proposed for retrieving/extracting informaion. The framework utilizes both the text and image in the process of obtaining information. The framework is general enough and can be applied for other information processing applications as well. This integrated view of looking at information will lead to better human-computer interaction. The framework can be modiifed by utilizing spatial realtionships between objects in an image to filter out the irrelevant results. We feel this will improve the results in some cases, e.g. medical images.

## References

1. Andreucci, A., Sneiders, E.: Automated question answering: review of the main approaches. In: ICITA '05 , pp. 514-519 (2005)
2. Barnard, K., Forsyth, D.:learning the semantics of words and pictures. In: International Conference on Computer Vision 2, 408–415 (2001)
3. Blei, D., Michael, Jordan, M. I. : Modeling annotated data. In: proceedings of 26th Annual international ACM SIGIR conference (2003)
4. Buitelaar, P., Sintek, M., Kiesel, M.: A lexicon model for multilingual/multimedia Ontologies. Proceedings of the third European Semantic Web Conference (2006)
5. Carson et al., 1997 Carson et al. (1997)
6. Hudelot, C., Maillot, N., Thonnat, M.: Symbol Grounding for Semantic Image Interpretation: From image data to semantics (2005)
7. Duygulu, P., Barnard, K., de Freitas, N. and Forsyth, D: Object recognition as machine translation: Learning a lexicon from a fixed image vocabulary. In seventh European Conference on Computer Vision, pages 97-112 (2002)
8. Jeon, J., Lavrenko, V, Manmatha Automatic Image annotation and retrieval using cross media relevance models. In the Proceedings of SIGIR'03, Toronto, Canada (2003)
9. Katz, B., Lin, J., Stauffer, C., Grimson, E.: Answering questions about moving objects in surveillance videos. In: Proc. of AAAI Spring Symposium on New Directions in QA (2003)
10. Boris, K.: (Annotating the World Wide Web using natural language', Proceedings of the 5th RIAO Conference on Computer Assisted Information Searching on the Internet (1997)
11. Ma and Manjunath 1998
12. Martin-Valdivia, M. T., Diaz-Galiano, M. C., Montejo-Raez, A., Urena-Lopez, L. A. Using information gain to improve multi-modal information retrieval systems, Information Processing Management 44, 1146–1158 (2008)
13. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In MISRM' 99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management. (1999)
14. Papadopoulou, G. T., Mezaris, V., Dasiopoulou, S and Kompatsiaris, I.: Semantic image analysis using a learning approach and spatial context. In Proceedings of the 1st International conference on semantic and digital media technologies (SAMT).(2006)

15. Su, L., Sharp, B., Chibelushi, C.: Knowledge-based image understanding: A rule-based production system for X-ray segmentation. In Proceedings of Fourth International Conference on Enterprise Information System, volume 1 (2002)
16. Town, C., Sinclair, D. Language-based querying of image collections on the basis of an extensible ontology. *Image vision Comput.* 22, 251–267 (2004)
17. Vompras, J.: Towards adaptive ontology-based image retrieval. In: Stefan Brass, C. G., editor, 17th GI-Workshop on the Foundations of Databases, Worfritz, Germany, pp.148–152. Institute of Computer Science, Martin-Luther-University Halle-Wittenberg. (2005).
18. Winograd, T.:*Understanding Natural Language*, Academic Press, New York (1973)
19. Yang, H., Chaisorn, L., Zhao, Y., Neo, S. Y., Chua, T. S Video QA: question answering on news video. In: Proc. Of ACM MM '03, pp. 632–641 (2003)
20. Yeh, T., Lee, J.J., Darell, T.: Photo-based Question Answering, ACM Multimedia (2008)
21. Möller, M. M., Sintek, M.: A Generic Framework for Semantic Medical Image Retrieval. In Proceedings of 7th Korea-Germany Joint Workshop on Advanced Medical Image Pro ( 2007)
22. Hudelot, C. Maillot, N. Thonnat, M.: Symbol Grounding for Semantic Image Interpretation: From Image Data to Semantics, In Proceedings of Tenth IEEE International Conference on Computer Vision (2005)
23. Siddiqui, T., Tiwary, U. *Natural Language Processing and Information Retrieval*. Oxford University Press. (2007)
24. Faraday, S A. Attending to Web Pages Pete Faraday, Microsoft, Redmond [http://www.cofc.edu/~learning/chi01\\_faraday.pdf](http://www.cofc.edu/~learning/chi01_faraday.pdf)

# Signal and Vision Processing

# An Application for Driver Drowsiness Identification based on Pupil Detection using IR Camera

K. S. Chidanand Kumar and Brojeshwar Bhowmick

Innovation Lab, Tata Consultancy Services Limited, Kolkata, India  
[{kschidanand.kumar,b.bhowmick}@tcs.com](mailto:{kschidanand.kumar,b.bhowmick}@tcs.com)

**Abstract:** A Driver drowsiness identification system has been proposed that generates alarms when driver falls asleep during driving. A number of different physical phenomena can be monitored and measured in order to detect drowsiness of driver in a vehicle. This paper presents a methodology for driver drowsiness identification using IR camera by detecting and tracking pupils. The face region is first determined first using euler number and template matching. Pupils are then located in the face region. In subsequent frames of video, pupils are tracked in order to find whether the eyes are open or closed. If eyes are closed for several consecutive frames then it is concluded that the driver is fatigued and alarm is generated.

## 1 Introduction

Road safety is discussed almost every day in the papers, airwaves in this country and worldwide. Unfortunately this is because of the lack of safety and precaution taken by road users. While dangerous driving may be highly publicized a major contributing factor in many accidents on our roads is driver fatigue, according to the National Roads Authority. The National Highway Traffic Safety Administration (NHTSA) [1] estimates that “1, 00,000 police-reported crashes are the direct fallout of driver fatigue each year”. This results in an estimated 1550 deaths, 71,000 injuries, and \$12.5 billion in monetary losses. In 2002, the National Sleep Foundation (NSF) reported [2] that 51% of adult drivers had driven a vehicle while feeling drowsy and 17% of them had actually fallen asleep. Recent statistics estimates that annually 1,200 deaths and 76,000 injuries can be attributed to fatigue related crashes.

Many papers [3-10], have been published to reduce automobile traffic accidents. However, best accuracy has been achieved by measuring physiological signals [10] such as brain waves, heart rate, and eye blinking. These techniques are intrusive, causing annoyance to drivers. A driver's state of vigilance can also be characterized by the behavior of the vehicle he/she operates. Vehicle behavior including speed, lateral position, turning angle, and moving course are good indicators of a driver's alertness level. While these techniques may be implemented non-intrusively, they are, nevertheless, subject to several limitations including the vehicle type, driver experiences, and driving conditions [3]. Fatigued people show certain visual behaviors like slow eyelid movement [5, 6], smaller degree of eye openness (or even closed), frequent nodding [7], yawning, gaze (narrowness in the line of sight),

sluggish in facial expression, and sagging posture. To make use of these visual cues, another increasingly popular and non-invasive approach for monitoring fatigue is to assess a driver's vigilance level through visual observation of his/her physical conditions using a camera and state-of-the-art technologies in computer vision. In a recent workshop [8] sponsored by the Department of Transportation (DOT) on driver's vigilance, it was concluded that computer vision represents the most promising non-invasive technology to monitor driver's vigilance. In [11], effort has been carried out using computer vision algorithm in order to detect pupil directly using differential method which subtracts the dark eye image (odd field) from the bright eye image (even field), producing a difference image, with most of the background and external illumination effects removed. Even though this method gives better results, this method relies more on hardware which is not compact and price wise it may be expensive. Despite the success of the existing approaches or systems for extracting characteristics of a driver using computer vision technologies, we focus on methodology which is of low cost, simple, accurate and has less response time.

The remainder of this paper is organized as follows. Section 2 briefly explains proposed methodology. Section 3 provides results and discussion. This is followed by conclusions and references.

## 2 Proposed Methodology

This paper focuses on the development of computer vision algorithms in order to extract information about opening or closing of eyes. In the following sub-section we first apply an adaptive threshold in order to remove background. The face region is located using euler number and template matching algorithm. Finally pupils are detected and tracked in consecutive frames.

### 2.1 Thresholding

The first step is to separate driver's face from background. Since IR camera focuses mainly on face, face is more illuminated than rest of the objects. Face region can be easily obtained by removing mean from input grayscale image  $I[i, j]$ , as shown in equation (1).

$$M[i, j] = I[i, j] - \bar{I} \quad (1)$$

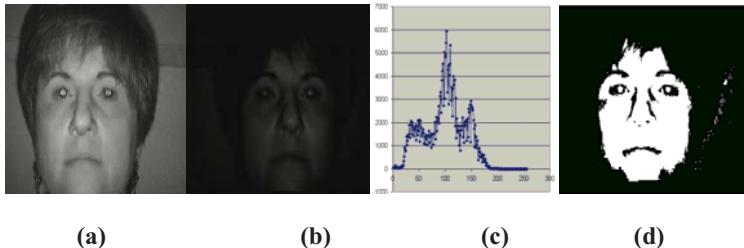
where M is Mean subtracted image, I is input grayscale image and

$$\bar{I} = \frac{1}{N} \sum_i \sum_j I[i, j], N \text{ is the area of the image.}$$

After obtaining mean removal image ( $M[i, j]$ ), the dynamic threshold is calculated using histogram technique. Figure 1(a) shows a typical input frame whose mean subtracted image is shown in Figure 1(b). The histogram for mean subtracted

image is plotted as shown in Figure 1(c). In order to detect threshold ( $T$ ), standard deviation ( $\sigma$ ) for the distance between all the peaks in histogram is calculated. If any adjacent peak falls within  $\pm 3\sigma$ , then that peak line is discarded. Now we search for a maximum peaks which is a face and it is considered as threshold ( $T$ ). If  $H[i, j]$  is the output image for an input image  $M[i, j]$ , then

$$\begin{aligned} H[i, j] &= 255 \text{ If } M[i, j] > T \\ &= 0 \text{ Otherwise.} \end{aligned}$$



**Fig. 1.** (a) Gray Scale Image. (b) Mean Subtracted Image (c) Histogram. (d) Binary Image

## 2.2 Face Detection

After thresholding, face and non face regions are distinguished. To detect the face region in the thresholded image, assuming that in the first frame of a video driver is not sleeping, i.e. eyes are opened. In Face region, eyes will come as holes and these holes play significant role to classify region as a face or non face. Euler number [13] is used to find the number of holes in a region.

The Euler number of an image is defined as

$$E = C - H$$

Where  $E$  is Euler number,  $C$  is the number of connected components and  $H$  is the number of holes in a region. As the face region has at least two holes for eyes, if any component having less than two holes then that component is rejected. Figure.2 shows only a face region excluding pupils.



**Fig. 2 Face region excluding pupils**

Now we have got the component where there are at least two holes in an image. But every component having two or more holes may not be a face. We use template matching algorithm in order to classify whether the component is a face or non face.

### 2.3 Template Matching

Template matching [14] uses a pre -defined model i.e. a human face. The model must be such that it rejects the non-face regions. The template face has to be positioned and rotated in the same coordinate system as the face image in order to have higher degree of matching. A typical template face model [14] is shown in Figure.3.



**Fig. 3. Template face model**

Template matching technique need to have knowledge of orientation and position of the region with which the template has to match. The position is characterized by the region centroid which is given by equation (2) and equation (3).

$$\bar{x} = \frac{1}{A} \sum_{i=1}^n \sum_{j=1}^m jB[i, j] \quad (2)$$

$$\bar{y} = \frac{1}{A} \sum_{i=1}^n \sum_{j=1}^m iB[i, j] \quad (3)$$

Where B is Matrix of size [n x m] of the blob, A is the area in pixels of the region.

Another important criterion is to find the orientation of the blob in order to have higher degree of matching with template model. The orientation of blob [14] is calculated as in equation (4)

$$\theta = 1/2 * \arctan \left( \frac{\beta}{\alpha - \lambda} \right) \quad (4)$$

Where:

$$\alpha = \sum_{i=1}^n \sum_{j=1}^m (x'_{ij})^2 B[i, j], \beta = 2 \sum_{i=n}^n \sum_{j=1}^m x'_{ij} y'_{ij} B[i, j], \lambda = \sum_{i=1}^n \sum_{j=1}^m (y'_{ij})^2 B[i, j],$$

$$x'_{ij} = x_{ij} - \bar{x} \text{ and } y'_{ij} = y_{ij} - \bar{y} .$$

Figure.4 shows a typical template matching scenario where Figure.3 is being imposed on Figure.1 (a). After calculating the position and orientation, we compute the cross-correlation between template face model and human face region to know the degree of matching between them. As per [14], a good threshold value for classifying a region as a face should be greater than or equal to 0.6.



**Fig. 4. Showing a typical template matching scenario**

## 2.4 Pupil Detection

A simple technique has been employed in order to obtain pupils from face region. Pupils can be obtained by performing XOR operation between thresholded binary image and euler number component binary image. This gives results of bright spots which indicate pupils, but it also contains stray images. In order to distinguish pupils from stray images, following rules are checked:

- 1) Axis of bright spots (pupils) must lies on same line.
- 2) Radius of bright spots (pupils) must have almost same area.

Figure.5 is resultant image obtained after performing XOR operation between Figure.1(d) and Figure.2. Some stray images that lie outside the bounding box of euler number component are filtered in this process.

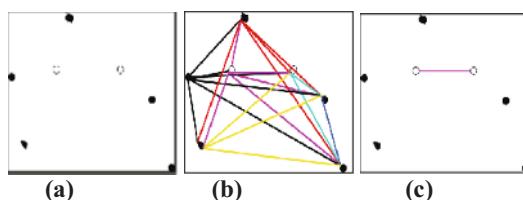


**Fig. 5. Resultant image**

## 2.5 Pupil Tracking

A simple technique called nearest neighborhood concept has been employed for pupil tracking which has reasonable accuracy and less computationally expensive. Once the pupils are detected, the left pupil area, right pupil area, their centroids & euclidean distance between these two centroids are calculated. A boundary box is created around these two pupils with  $\pm 10\%$  of width and height of an image. In the remaining frames of a video, search for pupil components whose area is approximately 2/3rd of right eye area or left eye area within the boundary box obtained from pupil detection part.

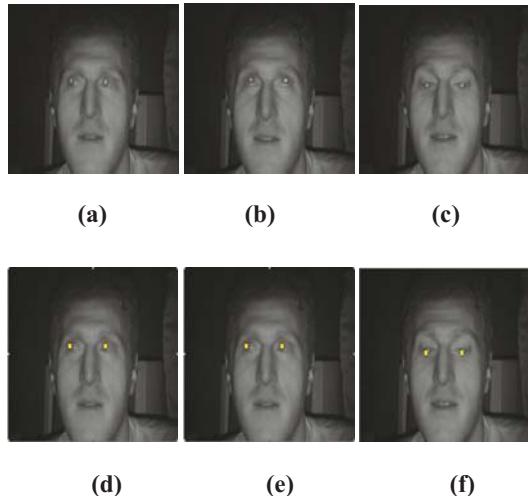
Consider an example as in Figure.6 (a), where black filled circles represent stray images and white circles represent pupils. Now compute euclidean distance between each centroid of circles, as shown in Figure.6 (b).



**Fig. 6. Distance between centroid of pupil**

The difference between pupil's centroids is always fixed, irrespective of whatever may be the position of driver. This indicates pupils as shown in Figure.6(c). This difference value is calculated in first frame of a video. When driver closes his eyes, then this distance will not be found in any component around the previous pupil positions and we conclude that the driver is sleeping. Figure.7 (d)-(f) shows how

pupils are tracked for video sequences in Figure.7 (a)-(c) using nearest neighborhood technique.



**Fig. 7.** Pupil Tracking

### 3 Results and Discussion

The proposed method was tested We tested with 5 videos [15] on 2.33GHZ PC, 1.95GB RAM.. The average correct rate of a video for driver drowsiness detection is 88.6%. The resolution of video is  $640 \times 480$  and  $720 \times 480$ . Figure.8 (a)-(e), shows various stages for detecting pupils in a non-sleeping video frame. Figure.8 (a) is an input video frame, Figure.8(b) is an image after subtracting mean from original image, Figure.8(c) is an image obtained after histogram thresholding process, Figure.8(d) contains face image excluding noisy part, Figure.8(e) is the result of XOR operation between Figure.8(f) and Figure.8(d), Figure.8(f) contains only pupils indicating driver is non-drowsy.

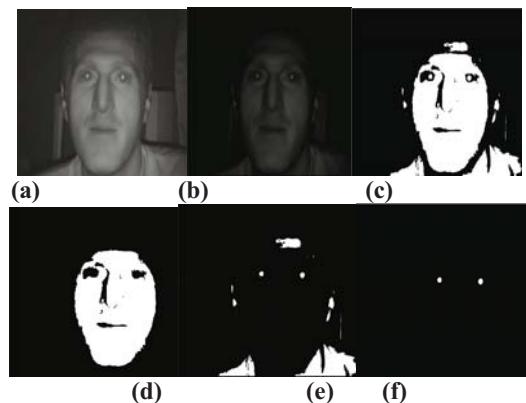
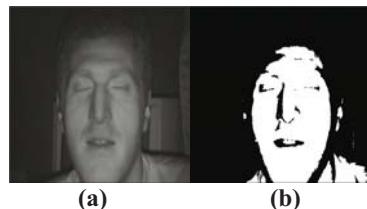
**Fig. 8.** Open Eyes Result

Figure.9 (a)-(b) show results for a sleeping video frame. In this video frame, pupils are directly tracked after thresholding process. In Figure.9 (b), pupils are tracked by pupil tracking algorithm as explained in previous section. Here there is no separate pupil component in a face and we will finally conclude that driver is sleeping in this frame.

**Fig. 9. Closed Eyes Result****Table.1** Result of driver drowsiness detection

	Video1	Video2	Vide3	Video4	Video5
Total Frames	92	375	92	92	92
Open Eyes	58	73	79	54	29
Closed Eyes	44	302	13	38	63
% Correct Rate	95	84	89	91	80
Average Correct Rate	88.6				

From the above Table, we observe that for video2 and video5 average correct rate is less. This is because video2 was taken during day time. During day time, daylight contains infrared light and it is difficult to get bright pupil effect which leads to false detection of pupils. In video5, average correct rate is lesser because of spectacles which create reflections that resemble pupils.

## 4 Conclusion

In this paper, we present driver drowsiness detection system based on face detection, pupil detection and pupil tracking methods. After finding the face, pupils are detected and tracked in order to determine opening or closure of eyes. Driver drowsiness is detected if the driver's eyes are closed for 5-6 consecutive frames. In the result section, for video2 and video5 average correct rate is lesser because of lesser bright pupil effect during day time and also because of reflections when driver wears spectacles. Future work is to reduce false alarms and make algorithms robust to reflections during daytime and also for driver wearing spectacles. Driver drowsiness detection technique can also be extended to other applications such as hand free interaction with computers (pointing and selection, application switching, scrolling, and document navigation) for disabled people based on pupil movement. Eye-tracking technology is reviewed along with new possibilities for measuring what athletes really see when they perform. It also helps sports trainer to improve athletes' attention, anticipation, concentration, memory, and problem-solving skills, leading to extraordinary long-term gains [16].

## 5 Acknowledgements

The author of this paper would like to thankfully acknowledge Mr. Aniruddha Sinha for his constant help to develop the entire system.

## References

1. NHTSA. Drowsy driver detection and warning system for commercial vehicle drivers (no date) "Field proportional test design, analysis and progress". National Highway TrafficSafety Administration, Washington
2. Weirwille, W.W: Overview of Research on Driver Drowsiness Definition and Driver Drowsiness Detection, 14th International Technical Conference on Enhanced Safety of Vehicles, pp 23-26 (1994)
3. Saito, H., Ishiwaka, T., Sakata, M.: Okabayashi S Applications of driver's line of sight to automobiles – what can driver's eye tell, Proceedings of 1994 Vehicle Navigation and Information Systems Conference, Yokohama, Japan, pp. 21–26 (1994)
4. Kaneda, M. et al.: Development of a drowsiness warning system. The 11th International Conference on Enhanced Safety of Vehicle, Munich (1994)

5. Daisi Onken, R., An adaptive knowledge- based driver monitoring and warning system, Proceedings of 1994 Vehicle Navigation and Information Systems Conference,Yokohama, Japan, pp. 3–10 (1994)
6. Feraric, J., Kopf, M., Onken, R: Statistical versus neural net approach for driverbehavior description and adaptive warning, The 11th European Annual Manual, pp. 429–436 (1992)
7. Ishii, T., Hirose, M. & Iwata, H Automatic recognition of driver's facial expression by image analysis". Journal of the Society of Automotive Engineers of Japan, 41: 1398–1403 (1987)
8. Hong Chung, K.: (Electroencephalographic study of drowsiness in simulated driving with sleep deprivation. International Journal of Industrial Ergonomics 35, 307–320 (2005)
9. Wierwille, W.W: Overview of research on driver drowsiness definition and driver drowsiness detection, ESV, Munich. (1994)
10. Dinges, D.F., Mallis, M., Maislin, G., Powell l.: Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. Department of Transportation Highway Safety Publication (1998)
11. Anon: Proximity array sensing system: head position monitor/metric, Advanced Safety Concepts, Inc., Santa Fe, NM87504.
12. Anon: Conference on Ocular Measures of Driver Alertness, Washington, DC (1999),
13. Thomson C.M, Shure L. Image Processing Toolbox, The Math Works Inc, USA.
14. Foo W., Chang H., Robles-MellinU.: Face Detection, Stanford University, Washington D.C, <http://www.csstudents.stanford.edu/~robles/ee368/main.html> (2000)
15. <http://www.humankinetics.com/products/showproduct.cfm?isbn=9780736042567>

# **Engine Fault Diagnosis using DTW, MFCC and FFT**

Vrijendra Singh and Narendra Meena

Indian Institute of Information Technology Allahabad, India  
vrij@iiita.ac.in, nrl@ug.iiita.ac.in

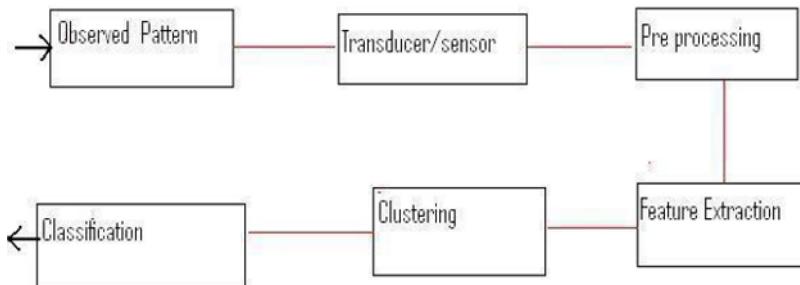
**Abstract.** Abstract. In this paper we have used a combination of three algorithms: Dynamic time warping (DTW) and the coefficients of Mel frequency Cepstrum (MFC) and Fast Fourier Transformation (FFT) for classifying various engine faults. Dynamic time warping and MFCC (Mel Frequency Cepstral Coefficients), FFT are used usually for automatic speech recognition purposes. This paper introduces DTW algorithm and the coefficients extracted from Mel Frequency Cepstrum, FFT for automatic fault detection and identification (FDI) of internal combustion engines for the first time. The objective of the current work was to develop a new intelligent system that should be able to predict the possible fault in a running engine at different-different workshops. We are doing this first time. Basically we took different-different samples of Engine fault and applied these algorithms, extracted features from it and used Fuzzy Rule Base approach for fault Classification.

## **1 Introduction**

The continuous monitoring and quality control of automobiles has motivated great interests in developing fault detection and isolation techniques. Generally faulty engine and faultless engine have different-different types of acoustics. Based on acoustics feature, a fault classification system has been developed. Pattern recognition method [1] can also be used for the same purposes.

DTW algorithm is most important and useful for fault diagnosis. DTW has been used by Choi et. al. [2], where along with DTW the process event detection algorithm has been used, the process event detection algorithm collects data through sensors and after that K-means clustering algorithm used for different-different classes and those different patterns are matched according to the pattern matching techniques of DTW. Generally, we use Cepstral coefficients as features for detecting fault. Use of Cepstral coefficients as feature has been used by Fulfilho et. al. [3], where they analyzed vibration signals of bearings for possible presence of bearing faults, here Cepstral coefficients of MFC were used with features extracted from MFD (Multi Scale Fractal Dimensions)and for classification they used HMM (Hidden Markov Model) and GMM (Gaussian Mixture Model). But this is the first time; we have used DTW along with MFCC and FFT for Engine fault classification. This paper has been organized

into the following sections at first we describe the different types of faults their nature, mode of origin, next we discuss the normalization and filtering of the



**Fig. 1.** Classical Pattern Recognition System

Engine audio samples, DTW, MFCC and FFT and Extracted Feature coefficients and their usage in classification of the faults.

## 2 The Different Engine Faults

We have diagnosed six types of faults, which incipient in nature in the beginning, might lead to serious deterioration of engine performance, they are shown as following:

### -TAPPET Fault

The rocker arm has a tappet as part and above the cylinder head of an internal combustion(IC) engine it make contact with an intake or exhaust valve stem. Space exists between the top of the valve stem and the rocker arm [4]. As the engine warms up, some mechanical expansion and lengthening of the valve stem and push rods allowed. If insufficient clearance is set when the engine is cold the valves will not properly close when the engine warms up. If too much clearance is provided then even after the engine warms up there will be some clearance which will result in lost motion. Over time mechanical wear causes an increase in clearance usually with the symptomatic ticking sound in the engine [5], called tappet noise. Here we take samples from two engines which have tappet fault, collected 8 samples, 4 from each.

### - CHAIN Fault

Cylinders in the automobile engines contain valves usually one intake and one exhaust. By using camshaft, valves are opened and closed. It is synchronized with the crankshaft so that it makes one revolution for every two revolutions of the crankshaft

[6]. In most engines, this is done by a chain and loss of tension in the chain, generates a grinding noise. This noise is Chain noise. Here we take samples from four engines which have chain noise, collected 16 samples, 4 from each.

#### **– Fault Related To CYLINDER CHAMBER**

The cylinder head sits at the top of the cylinders containing spark plugs in an IC engine and valves and a part of the combustion chamber. These valves are opened by a camshaft [7], defect in it causes a low humming noise, this is the main cause of CHN. CHN is two type, one is normal CHN and another one is Marginal CHN. We collected 4 samples of each.

#### **– Fault Related To KICK START CHAMBER**

It is a way of starting the IC engine, due to misplacement of the shaft the engine gives out a grinding noise during starting. This type of noise we called kick start noise. we collected 8 samples which have kick start fault.

#### **– FAULT Related To MAGNETIC CHAMBER**

The system includes a signaling coil, a delay circuit connected between the signaling coil and the thyristor switch. The delay circuit is reset every period of the output of the generating coil. As a result, the ignition timing is maintained substantially unchanged irrespective of the speed of the engine and electromagnetic noise from the signaling coil is greatly suppressed. We collected 4 samples from the IC-Engine which has Magnetic fault.

### **3 Experimental Setup and Data Collection**

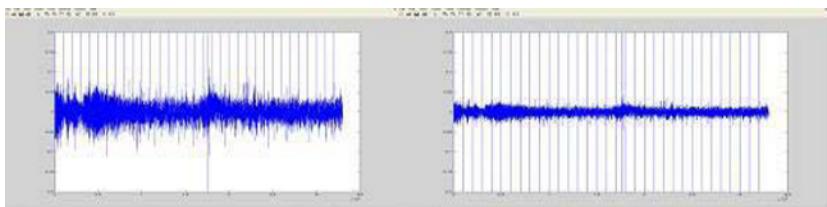
The experimental set up was carried out at different-different motorcycle work-shops of Allahabad city. A scheme was applied for proper data collection so that it best represents the fault. A microphone was positioned close to engine. The hardware used was NI DAQ and the software used was Lab view. ADAQ (data acquisition) was used for acquiring audio from the microphone; the sampling frequency was 100khz. Since, the black box figure of faulty engine arrival was very low; we just took sample of limited number of engines.

### **4 Normalization and Filtering of the Sample**

Normalization has been done for minimizing the variance in the signals. During recording of the faults various background noises like human voices were recorded with it. To remove the disturbances and extract the perfect feature for classification, pre emphasizing technique was used [9]. The signal was passed through a pre emphasis filter.

$$H(z) = 1 - a/z \quad (1)$$

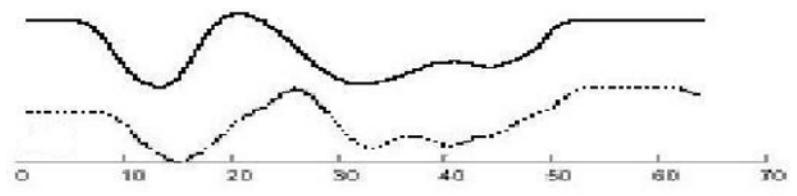
Where  $a = 0.95$ , the signal was sampled at 100Khz, the pre emphasized signal was windowed into overlapping segments of 25 ms with 10 ms before analysis. A tapered window function such as the Hamming window is used to reduce the effects of spectral leakage.



**Fig. 2.** Signal before and after Normalization and Pre- emphasis

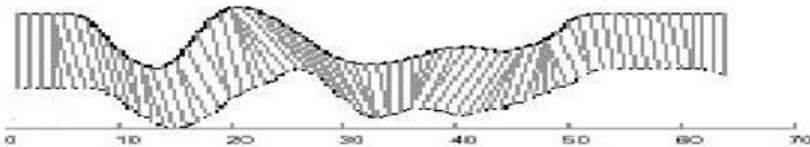
## 5 Dynamic Time Warping

Time series data is a very important problem in scientific experiments. A common task with time series data is comparing one sequence with another. In some domain the measurement of Euclidian distance suffices, but in cases where the two cases have approximately the same overall component shape, but they are not aligned in time axis, so to find the similarity between such patterns, we applied DTW using Dynamic Programming.



**Fig.3.** Two sequence having overall similar shape but not aligned

A proper explanation is given below:



**Fig. 4.** DTW can effectively find alignment between two sequence

### Step 1

After cutting and filtering of signal we get a filtered signal. We used spectrogram analysis of the signal.

### Step 2

– Now we use dynamic programming for alignment of signals (we can assume signals as a sequences). Suppose we took samples of tappet fault and we have to apply DTW. We collect 4 samples from Eng1 (Category#1) and another 4 from Eng2(Category#2).

- Each category have four sample (a0.dat, a1.dat, a2.dat, a3.dat) of audio Signals
- Now we want to find alignment of signals by using dynamic programming. First, we find alignment of a0.dat with a1.dat, a2.dat, a3.dat respectively. Again we find alignment of a1.dat with a2.dat, a3.dat respectively. And finally we find alignment of a2.dat with a3.dat.

For example, suppose the two (signal) sequences to be globally aligned are (sequence #1) and (sequence #2) at first the length of sequence #1 and sequence #2, are found. A simple scoring scheme is assumed where,

- $S(i, j) = 1$  if the residue at position  $i$  of sequence #1 is the same as the residue at position  $j$  of sequence #2 (match score); otherwise
- $S(i, j) = 0$  (mismatch score)
- $w = 0$  (gap penalty) So we had first initialized the matrix first column and rows by zero.

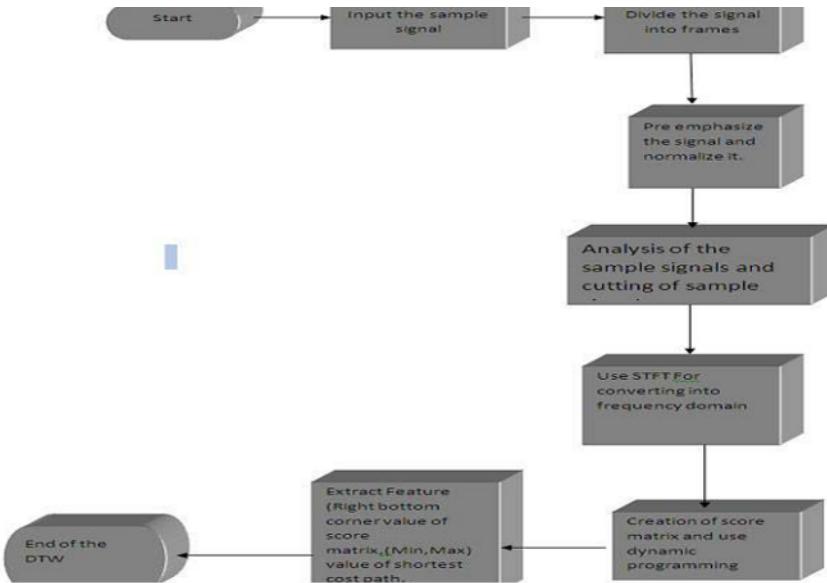
$$M(i, j) = \text{Max}[M(i - 1, j - 1) + S(i, j)(\text{match/mismatch in the diagonal}), M(i, j - 1) + w(\text{gap in sequence 1}), M(i - 1, j) + w(\text{gap in sequence 2})] \quad (2)$$

Note that in the example,  $M(i-1, j-1)$  will be red,  $M(i, j-1)$  will be green and  $M(i-1, j)$  will be blue. The score matrix is then filled and then the shortest cost path is calculated using trace back technique.

Features extracted from DTW will be described later.

## 6 FFT Implementation

The Fast Fourier Transform is an efficient procedure for computing the DFT (Discrete Fourier Transform) of a finite series and requires less number of computation. The FFT is based on decomposition and breaking the transforms into smaller transforms and combining them to get the total transform. After normalizing and filtering the signal, the signal is broken down into frames each typically lasting a fraction of a second. Then Hamming window is applied so that the edge effects can be reduced [11]. The application of hamming is done through Matlab (version 7.1) where the following equation is used to calculate the coefficients.



**Fig. 5.** Block Diagram of DTW

$$\omega(n) = 0.54 - 0.46\cos(2\pi \frac{n}{N}), \quad 0 \leq n \leq N \quad (3)$$

Then FFT is applied to each frame, the equation used for calculating FFT is given as

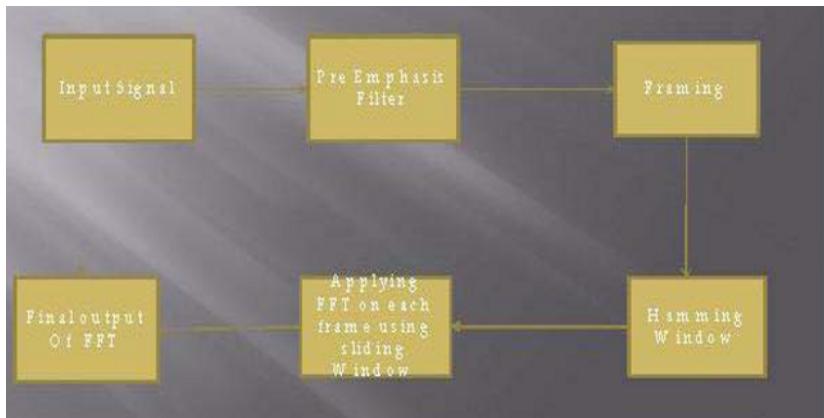
$$X(k) = \sum_{j=1}^N x(j) * \omega_N^{(j-1)(k-1)} \quad (4)$$

Where  $\omega_N = e^{(-2\pi i/N)}$ . Now by the application of sliding window the features are extracted.

## 7 MFCC

The Mel-frequency Cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a mel scale. Firstly, the whole signal is first divided into window of fixed size. Now, DFT on this window is computed for the discrete time signal  $x(n)$  of length N using the equation :

$$X(k) = \sum_{n=1}^{N-1} w(n) * x(n) * \exp(j * 2 * \pi * k * n / N) \quad (5)$$



**Fig. 6.** Block Diagram of FFT

Where  $k=0, 1, 2, N-1$ , where  $k$  corresponds to the frequency  $f(k) = kfs/N$ ,  $fs$  is the sampling frequency in Hertz and  $w(n)$  is a time-window. Here, we chose the popular Hamming window as a time window, due to computational simplicity. This part has already been described above in FFT. The magnitude spectrum  $|X(k)|$  is now scaled in both frequency and magnitude. First, the frequency is scaled logarithmically using the so-called Mel filter bank  $H(k, m)$  and then the logarithm is taken, giving

$$X(m) = \ln \left[ \sum_{n=0}^{N-1} |X(n)| * H(n, m) \right] \quad (6)$$

form = 1, 2, . . . , M, where M is the number of filter banks. The Mel filter bank is a collection of triangular filters defined by the center frequencies fc (m), written as:

$$\begin{aligned} H(k, m) &= 0, \text{ for } f(k) < fc(m - 1) \\ &= \frac{f(k) - fc(m - 1)}{fc(m) - fc(m - 1)}, \text{ for } fc(m - 1) - f(k) < fc(m) \\ &= \frac{f(k) - fc(m + 1)}{fc(m) - fc(m + 1)}, \text{ for } fc(m) - f(k) < fc(m + 1) \\ &= 0, \text{ for } f(k) - fc(m + 1) \end{aligned} \quad (7)$$

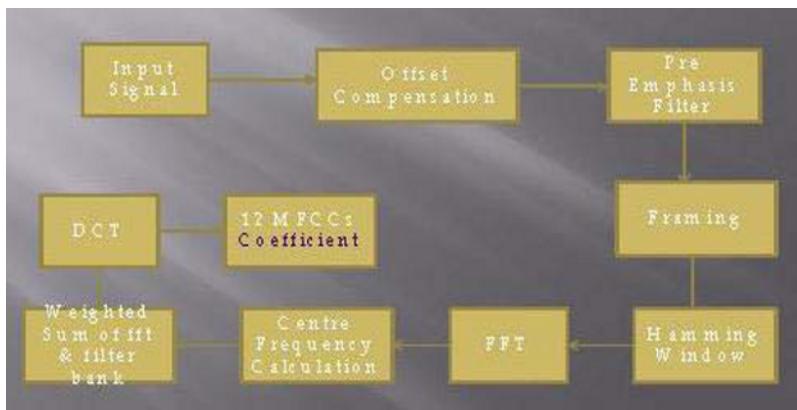
Here, the centre frequencies of filter bank are computed by approximating the mel scale with  $Si = 2595 \log(f/700+1)$ . This is a common approximation. A fixed frequency resolution in the Mel scale is computed, corresponding to a logarithmic scaling of the repetition frequency, using  $\delta Si = (Si_{max} - Si_{min})/(M+1)$ , where  $Si_{max}$  is the highest frequency of the filter bank on the Mel scale, computed from  $f_{max}$  using equation (4),  $Si_{min}$  is the lowest frequency in Mel scale, having a corresponding  $f_{min}$ , and M is the number of filter banks. The center frequencies on the Mel scale are given by  $Sic(m) = m * \delta Si$  for  $m = 1, 2, \dots, M$ . To obtain the center frequencies in Hertz, we apply the inverse of equation (4), given by

$$fc(m) = 700(10^{(\delta Sic(m)/2595)} - 1) \quad (8)$$

which are used to give the Mel filter bank. Finally, the MFCCs are obtained by computing the DCT of  $X'(m)$  using

$$c(l) = \sum_{m=1}^M X'(m) * \cos((1 - \pi(m - 1/2))/M) \quad (9)$$

for  $l = 1, 2, \dots, M$ , where  $c(l)$  is the lth MFC. Features extracted from MFCC will be described later.



**Fig. 7. Block Diagram of MFCC**

## 8 Feature Extraction and Fault Analysis Model

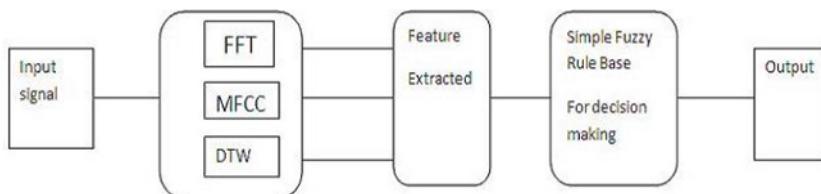
Three algorithms DTW, MFCC and FFT are applied on the different lasses(as mentioned in section 2), and the corresponding features for that particular class is extracted, now the different features that we have considered for the above mentioned algorithms are stated below.

DTW, first we considered the value of right bottom corner value of the score matrix, the maximum and minimum values of shortest cost path and all the diagonal element of score matrix. FFT, we considered maximum and minimum frequency of each frame and after that we considered the maximum and minimum of the entire frame. MFCC, we considered the maximum and minimum values of cepstral coefficients. We applied MFCC, FFT, DTW simultaneously on each samples of each fault and did analysis of all feature value. All Feature values, which we used for making Rule base has been given in table 1.

**Table 1.** Feature Extraction table

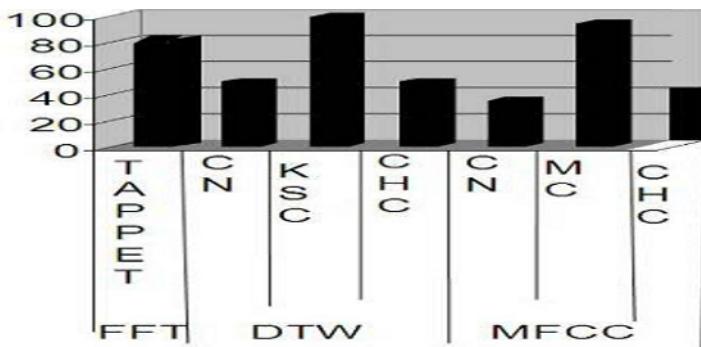
Feature Extracted Values	Algorithm Applied	Fault to be classified
30.8230	MFCC	Fault related to Magnetic chamber
30.8484		
30.8752		
30.7067		
30.6295	MFCC	Fault related to Cylinder Chamber
30.6711		
30.6074		
30.6270		
30.6345	MFCC	Chain Noise
30.6954		
30.6573		
30.6505		
30.6044	DTW	Fault related to Cylinder chamber
30.6295		
30.6711		
30.6074		
30.6270	DTW	Fault related to Kick Start chamber
5.520		
65		
1.0479		
1.8670	DTW	Chain Noise
2.1439		
5.0979		
30.5536		
30.9053	DTW	Tappet Noise
0.0258		
0.0270		
0.0333		
0.341	FFT	
0.0311		
0.313		
0.0269		
0.0299		
0.0553		
0.0501		
0.0371		
0.291		

After the feature extraction procedure an empirical fault Analysis model has been shown in the following figure.

**Fig. 8.** The Fault Analysis Model set up

## 9 Conclusions

After the empirical Fault Analysis Model was setup it was tested for signal belonging to any class and observations has been shown in the following histogram:



**Fig. 9.** A Histogram showing the percentage classification by FFT, DTW and MFCC of each fault classes

In figure:

CN: Chain noise

CHC: Cylinder head chamber

MC: Magnetic chamber

KSC: Kick start chamber

FFT could only classify Tappet fault, MFCC has been able to classify three fault types. Thus MFCC and DTW usually use for speech recognition process has been able to classify engine faults. So combining with FFT this method can be successfully used for automatic engine fault recognition system, saving maintenance cost of engine by indicating timely repair and precious lives.

## 10 Future Work

Future work is being targeted by taking consideration that one or more faults can occur in an engine at the same time.

## Acknowledgments

We are grateful to TIFAC, Govt. of India, for providing us the research grant

## References

1. J. C. C., Castejn O. :Incipient bearing fault diagnosis using DWT for feature extraction, World Congress, Besanon (France).
2. Ja Choi, Jong Myoung Ko, Chang Ouk Kim, Yoon Seong Kang, Seung Jun Lee.: Dynamic Time Warping Algorithms for Run-by-Run Process Fault Detection, ISSM Paper: PC-P-101
3. Fulufhelo V., Tshilidzi M., Unathi M.: Early classification of bearing faults using Hidden Markov Models, International Journal of Innovative Computing, Information and Control.
4. Wen-Xian Y.: Diagnosing the engine valve faults by genetic programming, Journal of Sound and Vibration Volume 293, Issues 1-2. 5. <http://www.motorera.com/dictionary>
6. Ren Y., Hu Tianyou, Yang P., Liu X.: Approach to diesel engine fault diagnosis, Mechatronics and Automation, 2005 IEEE International Conference 3, 1451–1454 (2005)
7. Ren Y., Hu Tianyou; Yang P. Liu X.: Approach to diesel engine faultdiagnosis, Mechatronics and Automation, IEEE International Conference Volume 3, 1451–1454 (2005)
8. Shiyuan L., Fengshou G., Ball A.: Detection of engine valve faults by vibration signals measured on the cylinder head, Journal of Automobile Engineering.
9. Kobayashi, K.; Uchikawa, Y.; Simizu, T.; Nakai, K.; Yoshizawa, M. :The rejection of magnetic noise from the wire using ICA for magneto cardiogram, Magnetics, IEEE Transactions 41(10), 4152–4154 (2005)
10. Ciaramella, E.; Giorgi, L.; Dapos;Errico, A.; Cavaliere, F.; Gaimari, G.; Prati, G. “Technique for setting the power preemphasis in WDM optical systems”, Journal of Light wave Technology, 24(1), 342–356 (2006)
11. Srinivasan R., Ming S.: Online fault diagnosis and state identification, Chemical engineering science
12. Sherlock B. G. and Kakad Y. P. :Transform domain technique for windowing the DCT and DST, Journal of the Franklin Institute 339(1), 111–120 (2002)
13. Duan C., He Zhengjia, and Jiang H.: A sliding window feature extraction method for rotating machine based the lifting scheme, Journal of Sound and Vibration Volume 299, Issues 4–5

# Multimodal News Story Segmentation

Gert-Jan Poulsse and Marie-Francine Moens

Katholieke Universiteit Leuven, Department of Computer Science, Celestijnenlaan  
200A Box 2402  
B-3001 Heverlee, Belgium  
{Gert-Jan.Poulsse, Marie-Francine.Moens}@cs.kuleuven.be

**Abstract.** In this paper, we describe a multi-modal approach to segmenting news video based on the perceived shift in content. We divide up a video document into logically coherent semantic units known as stories. We investigate the effectiveness of a number of multimedia features which serve as potential indicators of a story boundary. The results show an improvement of performance over current state of the art story segmenters.

## 1 Introduction

The aim of our research is to implement accurate methods for story segmentation in news video. In this context, this means detecting the specific time event at which one news story stops being discussed and a new story starts being reported. In text, a story is a coherent grouping of sentences, discussing related topics and names. The multimedia equivalent, such as found in news video, would be a temporal segment containing imagery accompanied by a spoken description of the single news event.

Three different channels, text, video, and audio are at our disposal for the segmentation task. Our aim is to base the segmentation decision on the detected change in content across the various media. Although considerable work has been done in developing story segmenters that utilize numerous multimodal features we would like to investigate some of the text based features and methods developed in research to date. We wonder whether combining the various approaches into a single, unified segmentation algorithm might not improve performance of segmenting broadcast news video. In order to effectively operate on this multimodal domain, we also include video and audio features in our investigation. Since our segmentation results form a basis for additional tasks such as summarization and concept detection, we wish to obtain the lowest possible error rate and so we introduce supervision to our segmentation efforts. In order to do this, we train a maximum entropy classifier on various multimedia features.

## 2 Related Work

Initial efforts at topic segmentation in text determine the lexical cohesion by measuring vocabulary repetition, as expressed by the cosine score of the term vectors representing two adjacent blocks of text. (Hearst 1997) assigns a story break between text blocks whose cosine scores differ greatly. (Choi 2000) computes an inter-sentence ranking from the cosine. Story segments are then identified by maximizing this ranking score while recursively partitioning the text. Other approaches use language models (Beeferman et al, 1999) or lexical chains (Kan et al. 1998, Galley et al. 2003) to compute lexical cohesiveness.

Segmentation of spoken discourse includes work done by (Passonneau et al. 1993, Galley et al., 2003) and makes use of a number of indicators such as cue-words, pause duration, and other forms of speech prosody. Work done for the TRECVID 2004<sup>1</sup> story segmentation task (of news video) is noteworthy as the approaches taken are more grounded in video retrieval. A representative example is IBM (Amir et al. 2004), who combine numerous visual features with specialized commercial and anchor (news reader) detectors, speech prosody, and textual features in order to find story boundaries.

## 3 Multimedia Features

Our intent is to identify story boundaries, using sentences as the candidate points between which story breaks occur. The following sections describe the features extracted from a multimedia document, and the motivation behind their choice. Ultimately these features will be used to train a maximum entropy classifier which will determine the existence of a story break at a particular sentence.<sup>2</sup>

**Segment Likelihood.** When considering whether to place a boundary at a candidate point, one can gauge the effect of preserving the segment integrity, versus splitting it up into two new segments, by computing the difference of the likelihoods that words within a segment are generated from the original segment or the two new segments.

$$\text{Score}(i) = \frac{L(\text{original})}{L(\text{original}) + L(\text{new segments})}$$

$$L(\text{segment}) = \prod_{\text{words}} L(\text{word}|\text{segment}) = \prod_{\text{words}} \alpha P(\text{word}|\text{segment}) + (1-\alpha)P(\text{word}|\text{wiki})$$

<sup>1</sup> <http://www-nlpir.nist.gov/projects/tv2004/tv2004.html>

<sup>2</sup> For the sake of brevity, we omit discussions of features related to vocabulary repetition as measured by the cosine metric as implemented by (Hearst 1997) and (Choi 2000), topic similarity as determined by Latent Dirichlet Allocation (Blei 2003), and news broadcast program structure, as these features did not contribute to the final solution.

$$P(\text{word}|\text{segment}) = \frac{\#\text{word occurrences in segment} - 1}{\text{total } \# \text{ words in segment} - 1}$$

The likelihood function measures term repetition within a segment smoothed by the chance of it occurring naturally, as defined by term frequencies gathered from a large external corpus, which in our case was Wikipedia, which because of its diversity we consider it to be topic neutral. The resultant score is used as a feature.

**Story Size.** A layout related feature that we consider is the story size of the previous segment. The reasoning behind this is that the highlights section of a news broadcast consists of many, short consecutive passages. Thus the presence of short story segment, corresponding to such a story highlight, might indicate that another short segment might soon follow. This feature is certainly domain driven, but not entirely inconceivable.

**Speech Pauses.** Work done by among others (Passonneau 1993) has shown that speech prosody can contribute to the detection of story breaks, with speaker pause duration often being the most important feature. We assume a larger pause between stories than between the sentences of a story. We therefore used a voice-activity detector (Dekens et al. 2007) to extract the duration of all the silences in the audio channel. We identified the longest silence fragment immediately preceding a sentence, and used this duration as a feature.

**Shot Cuts.** The rapid change in visual content, usually due to some camera motion or change in scenery is referred to as a shot cut. One supposes that a visual change would be correlated with a change in semantic content. Like (Amir et al. 2004), we identify visual transitions using a shot cut detector provided by (Osian and van Gool 2004). Unfortunately, shot cuts do not necessarily indicate a story boundary. There are generally many shot cuts within a single story unit. Often shot cuts do not always precisely align with sentence boundaries. As a result we computed the shot cut feature by examining each sentence in the document and assigning a binary feature, which indicated the presence or absence of a shot cut during the time period in which a sentence was uttered. This time window was slightly offset, in order to catch shot cuts that occurred immediately prior to, or after, the sentence was read.

**Cue Words: Chi-Square.** In many works, such as (Beeferman 1999) and (Galley et al., 2003), cue words and phrases are used as a feature for the detection of topic breaks when segmenting text. Phrases such as “good morning,” or “this is Alastair Yates,” are commonly said by news anchors or reporters to begin or end a particular news story, and their detection might help in recognizing story breaks. We examined the phrases immediately preceding or following a story break in a training set, and compared this with how often they occurred in the rest of the corpus. We performed  $\chi^2$  tests at a significance level of 0.01 in order to determine the phrases indicative of a story transition. We discovered such phrases as, “Hello

and welcome to BBC news”, “Good evening”, “Stay with us”, and “These are the headlines.” When performing feature extraction, each sentence receives a binary score indicating the presence or absence of cue phrases.

**Implicit Cue Words.** We implicitly learned cue phrases by training a maximum entropy classifier on sentences in our training set which were on, or away from, story boundaries. The probability score returned by this classifier forms one of our features.

**Named Entity Chains.** (Hearst 1997, Kan et al. 1998, Galley et al. 2003) use lexical chains to measure entity repetition and thus infer the cohesiveness of a prospective story segment. We observed that news stories often have disjoint sets of named entities, i.e. the proper names of people, places, or organizations. Applying a limited form of lexical chaining of named entities, the density of these chains over a segment closely mirrors the actual story segment; boundaries occur where there are few chains.

It should be noted that longer interviews do not as closely follow this pattern as there generally are less named entities due to dialogue etc. Also, very short story segments, such as the highlights section of a news broadcast, cannot be distinguished using this method as they generally have a similar amount of named entities.

We used the Stanford Named Entity Recognizer (Finkel et al, 2005). We calculated the number of named entity chains spanning every sentence as a feature.

**Lexical Chains: Galley.** We include the score feature developed by (Galley et al. 2003), which measures based on lexical chains. He combines the frequency of term repetition with chain compactness to arrive at a descriptor for lexical cohesiveness. This is then used to compute the rate of change in lexical cohesiveness, where a low score indicates a story boundary.

## 4 Story Boundary Selection

The number of story segments to find in a document can either be specified or be estimated from the training data. We adopt the following heuristic, we determine the ratio of sentences to story segments from the training data, and apply this figure to determine the number of segments in unseen documents in our test set.

After a random initialization, where we place that amount of boundaries in the document, we use an iterative method to reassign story boundaries based on a fitness criterion-a maximum entropy classifier trained using features from section 3. Purported story segments with a low fitness score will be removed and new story boundaries will be placed at random position elsewhere. Over many iterations, certain candidate boundary positions (sentences) will more often have a

story boundary occur on them. The candidate boundary positions for which this occurs most frequently are returned by the algorithm.

## 5 Evaluation

We have collected and annotated 14 news broadcasts from the BBC, which have a combined duration of around 7 hours. In addition, we have the corresponding transcripts for each broadcast, which consist of over 3000 sentences in total. The transcripts were sometimes noisy due to transcription errors, and an effort was made to correct the worst distortions by hand.

In text based segmentation, quite a few evaluation metrics have been proposed. Early papers such as (Hearst 1994) used the precision/recall metrics, although this metric is too strict in that it penalizes boundaries that have been placed very close to, but not on the ground-truth boundary. As a result, degenerate algorithms which place a boundary after every possible sentence can actually achieve a higher precision and recall score. (Beeferman et al. 1999) proposed a metric,  $P_k$ , which penalizes degenerate algorithms yet also gives partial credit for boundaries which are close to the actual boundary. An improvement on the  $P_k$  metric, called WindowDiff (WD), was introduced by (Pevzner and Hearst 2001).

We performed leave-one-out (leaving out 1 broadcast every time) cross-validation to train and evaluate our segmentation algorithm. The average  $P_k$  and WD scores were computed from the held-out test sets.

We established a strong baseline, by using our likelihood function for lexical cohesion, story segment size, and automatically learned cue phrases. We then incrementally added in the remaining features into our maximum entropy classifier. We then considered the effects of a late fusion of our best performing maximum entropy classifier with the remaining unused features. Late fusion was performed as follows:

$$P(\text{story break} \mid \text{Classifier}) = \lambda_{i+1} P(\text{story break} \mid \text{MaxEnt}_{\text{best}}) + \lambda_{i+1} P(\text{story break} \mid \text{feature}) + \dots$$

A new classifier was trained on our training set by interpolating the results of our best Maximum Entropy classifiers with our unused features. Interpolation weights were determined with the use of the Expectation Maximization (EM) algorithm. Only cue words and the lexical chain (GAL) feature gave an improvement.

We list the runs which showed a consistent performance increase in Table 1.

**Table 1.** Best performing feature combinations.

		BASE +NE+ GAL	BASE+ PAUSE + CUE+ GAL	BASE+NE+GAL+ Late Fusion(CUE)	BASE+PAUSE+ CUE+GAL+ Late Fusion(CUE)	BASE+PAUSE + CUE+GAL+La te Fusion (CUE+GAL)
WD	0.212	0.209	0.210	<b>0.194</b>	0.198	0.197
P <sub>k</sub>	0.135	0.142	0.141	<b>0.119</b>	0.122	0.121
BASE is the combination of segment likelihood, segment size, and implicit cue words						
PAUSE is the pause duration feature				Cue is the $\chi^2$ Cue Words		
NE is Named Entity chain feature				GAL is the score defined by Galley		

We then evaluated our resultant system against Galley's segmenter (Galley et al., 2003), LCSeg, and Choi's segmenter (Choi 2000), C99. These are two state of the art systems which segment text in an unsupervised fashion based on text-only features. Our system differs from theirs in that we use multiple features, including those specific to multimodal data. Both C99 and LCSeg, like ours, are capable of automatically determining the number of segments in a document. We provide results for both modes of operation, referred to as *known* (number of segments is known) and *unknown* respectively in Table 2. Both C99 and LCSeg ran using their default parameters.

**Table 2.** Comparison of the results from C99 and LCSeg.

	BASE+NE+G AL+Late Fusion (CUE) known	BASE+NE+G AL+ Late Fusion (CUE) <sub>unknown</sub>	C99 <sub>known</sub>	C99 unknown	LCSseg known	LCSeg <sub>unknown</sub>
WD	<b>0.194</b>	0.220	0.363	0.307	0.276	0.243
P <sub>k</sub>	<b>0.119</b>	0.149	0.323	0.268	0.218	0.191

## 6 Conclusion

Our initial belief that a unification of several features and methods from the textual modality would result in improved segmentation performance was validated by our final result. The inclusion of multimedia features, such as shot cuts and pause duration, only resulted in a performance increase with the pause feature. Data analysis revealed that too many shot cuts occurred during a single story segment, both on actual boundaries and on intra-segment sentences, meaning this feature was insufficiently discriminative. Ironically our best segmentation of our news video dataset used only features from the textual modality.

We acknowledge that many previous segmentation efforts in research have focused on unsupervised methods, yet we also feel we have more than adequately

demonstrated the improved performance made possible by the use of a supervised training method. In light of the small development set requirement (13 broadcasts), and future mission critical applications (document summarization and concept detection), this seems a justified choice, as the performance of our segmenter exceeds that of other methods.

## References

1. Amir, D., Argillander, Berg, J. M., Chang, S-F.: IBM Research TRECVID-2004 Video Retrieval System, Proc. TRECVID (2004)
2. Beeferman, D., Berger, A. J. Lafferty: Statistical models for text segmentation Machine Learning 34 ( 1-3), 177–210 (1999)
3. Blei,D. M., Ng, A. Y., Jordan. M. I.: Latent Dirichlet Allocation. Journal of Machine Learning Research, 993–1022 (2003)
4. Choi, F.: Advances in domain independent linear text segmentation. Proc NAACL (2000)
5. Dekens, T., Demol, Verhelst, M., Beaugendre W. F.: Voice Activity Detection based on Inverse Normalized Noise Likelihood Estimation. CIE 2007, Santa Clara, Cuba, June 18–22 (2007)
6. Finkel, J., Grenager, T., Manning, C.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. ACL (2005)
7. Galley, M., McKeown, K., Fosler-Lussier, Jing, E. H.: Discourse Segmentation of Multi-party Conversation. Proc. ACL. Sapporo, Japan,pp 562-569 (2003)
8. Hearst, M. A.: TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. Computational Linguistics, 23(1), 33–64 (1997)
9. Kan, M-Y., Klavans, J. L., McKeown. K. R.: Linear Segmentation and Segment Significance." Proc. 6th Workshop on Very Large Corporation (1998)
10. Osian, M., Van Gool, L.: Video shot characterization. Machine Vision and Applications, 15(3), 172–177 (2004)
11. Passonneau, R. J., Litman D. J.: Intention-Based Segmentation: Human Reliability and Correlation with Linguistic Cues. Proc. ACL, pp. 148–155 (1993)

# **RGB Color Histogram Feature based Image Classification: An Application of Rough Reasoning**

Shailendra Singh

Samsung India Software Center, Noida, India

Email: {shailendra.s@samsung.com, dr.shail@gmail.com}

**Abstract.** In this paper, we have proposed a novel rough set based image classification method which uses RGB color histogram as features to classify images of different themes. We have used the concept of discernibility to analyze RGB color values and finding optimum color intervals to discern images of different themes. We have shown how the set of optimum color intervals for training set of images can be used to construct a representative sieve for training set of images to classify new set of images. We have also proposed rough membership based formulation to classify new set of images using representative sieve. The outlines of the system has been implemented and presented along with some results on a set of new images of different themes.

## **1 Introduction**

In today's world, the use of digital images is becoming very popular. As of result, we have very large collection of images almost on each domain. These collections may have natural scenes, historical monuments, plants, mountains, animals, buildings, medical images, and many more. Therefore, we can simply conclude that how important images in our daily life. Also it is becoming difficult to find specific images in the collection of images. Researchers are finding different ways to manage collection of images. The simplest way is based on text query where user enters text query and search relevant images based on their meta-tag information but results are not accurate. Therefore researchers shifted their focus on Content based image retrieval and classification [4]. Some of the approach analyzes content of an image and retrieves similar images from large image repository while other approaches classifying collection of images into groups of similar images.

In this paper, we have explored the possibilities of applying rough-set based reasoning for image classification. Since images cannot be uniquely categorized, rough categorization is ideally suited for the problem of finding relevant image features. The main contributions of this paper are

- Handling the problem of choosing relevant RGB color histogram based optimum color sub-intervals to discern between images of different themes. We proposed the concept of discernibility to a set of training images which is having images from " $n$ " different themes. The set of *optimum color intervals*

obtained through this process is used to construct a representative sieve which is used to classify new set of images. It has been found that the discernibility reduces dimensionality of image features in the form of representative sieve which reduces the computation time in classification of new set of images.

- The usual rough membership functions use overlaps of equivalence classes in terms of a subset of features to find membership of an element to different decision classes. This process cannot be applied straight away for image classification since an image decision class (or theme) usually cannot be associated to a particular value of pixels in color sub-intervals. We have proposed a rough membership based classification method to classify new set of images. We have provided performance analysis of the classification method by comparing the system generated ratings of the images with user ratings.

The rest of the paper is organized as follows. In section 2, we have reviewed some ongoing work in the area of image classification as well as rough-sets based approaches in the field of imaging. In sections 3, 4, and 5, we have provided the rough-theoretic analysis of our learning system. In section 6, we have presented a new rough membership computation method to classify images. In section 7, we have provided performance analysis and conclusion of our proposed system.

## 2 Related Work in Image Classification

Image classification is a very challenging research area. Researchers are continuously finding different ways and different approaches for better results. In [1] chapelle proposes Histogram-based image classification using Support Vector Machines. This paper shows that support vector machines (SVM's) can generalize well on difficult image classification problems where the only features are high dimensional histograms. They have used heavy-tailed RBF kernels and evaluated on Corel stock photo collection. Wang [14] proposes a new color feature representation which not only takes the correlation among neighbouring components of the conventional color histogram into account but removes the redundant information and uses SVM and Adaboost. Sergyan [9] proposes the possibility of image classification using certain color descriptors and the usage of different color spaces. [11] proposes a color image retrieval method based on the primitives of images. They consider context of each pixel in an image then clustered into several classes based on the algorithm of fast non-iterative clustering

Park [7] proposes a method of content-based image classification using a neural network. The classification is divided into foreground and background object images and uses neural network based back-propagation learning algorithm classifier is constructed for classification. Vailaya [13] are capturing high-level concepts from low-level image features using binary Bayesian classifiers. They are classifying images as indoor or outdoor; outdoor images are further classified as city or landscape; finally, a subset of landscape images is classified into sunset, forest, and mountain classes. Mrowka [5] proposes an approach using a non parametric statistical test for effective dimensionality reduction.

Rough set introduced in 1980 by Prof. Pawlak [8]. This approach is gaining popularity in diverse research areas especially in knowledge discovery and artificial intelligence. Rough sets have also been used in imaging filed, however, the application of rough sets for color image analysis has yet to be fully investigated. Mohabey [6] proposes a new concept of encrustation of the histogram, called histon, has been proposed for the visualization of multi-dimensional color information in an integrated fashion and its applicability in boundary region analysis using rough set based approximations. Dong [2] proposes remote sensing image classification algorithm based on rough set theory. Yun [15] proposes a rough neural network (RNN) to classify digital mammography to detect tumor. The experimental results show that the RNN performs better than purely using neural network in terms of time and classification accuracy. Shang [10] presents fuzzy-rough based feature selection and neural network-based classification approach. Singh [12] proposes dimensionality reduction method using discernibility approach and further using rough membership based ranking method for web documents.

### **3 Rough Set Based Methodologies to Determine the Most Discerning RGB Color Parameters**

A rough set [8] is an information system which is used for classificatory analysis. It is particularly useful when the elements of the domain do not lend them to a unique classification. An information system can be defined as a pair  $I = (U, S)$ , where  $U$  is a non-empty finite set of objects called the universe and  $S$  is a non-empty finite set of attributes. An information system is called a *decision system*  $D$  if it has an additional decision attribute  $d$ . This is represented as,

$$D = (U, S \cup \{d\}), \text{ where } d \notin S \text{ is the decision attribute} \quad (1)$$

At the core of all rough set based analysis lies an equivalence relation called the *indiscernibility relation*. For any  $A \subseteq S$ , the equivalence relation,  $\text{IND}_I(A)$  is defined as

$$\text{IND}_I(A) = \left\{ (x, x') \in U^2 \mid \forall s(x) \in A, s(x) = s(x') \right\} \quad (2)$$

#### **3.1 Determine the Most Discerning RGB Color Parameters**

In this section, we will illustrate how the most discerning RGB color parameters can be obtained using discernibility concept as explained in [3] and [12] by analyzing a set of training images from different classes. Let  $x$  be an *information item* to be retrieved from a finite set of items  $U$ . In our case  $x$  represents an image. An *information feature* is an attribute used for retrieving an image. For image retrieval, *information feature* can be RGB color intervals, CMY colors, HSV

colors, and other image related features. In our approach, we are using RGB color intervals as *information features*.

Each image  $x_i$  in our universe is represented as a set of RGB colors and each color can be defined in the range of [0-255]. In our proposed approach, we divide each color range in to 8 sub-intervals like [0-31], [32- 63].....[224-255]. Therefore, we will have 24 sub-intervals for red, green, and blue colors as information features. Our aim is to define each image in the form of sub-intervals and their weighted value which will be represented by no. of pixels occurring in that sub-interval. With this concept, we build our initial decision system for classification of images. We collect a set of training images retrieved from a standard image search engine like Google image search. This training set of images can be classified into “n” no. of decision classes. We convert each training image in its weighted value representation as described above. A decision table is shown in equation (3) where  $n w_{ij}$  represents the weight of information feature  $w_j$  in information item  $x_i$ .

$$D_T = (W_M \cup \{d_i\}) \text{ where } 1 \leq i \leq n,$$

$d_i$  is the decision class for a training image  $x_i$ ,

$$\text{and } W_M = [n w_{ij}]_{n \times k} \text{ where } 1 \leq i \leq n, 1 \leq j \leq k$$
(3)

Since the values stored in the decision table are continuous, it is not ideally suited for finding optimum color sub-intervals. Thus we first *discretize* the table. The discretization technique is as follows.

Let the range of values corresponding to any information feature (sub-intervals) be  $V_s = [I_s, L_s] \subset \mathfrak{R}$ . A partition  $P_s$  is induced on  $V_s$  by considering all values occurring for the sub-interval  $s$ , and arranging them in ascending order. Thus

$$P_s = \{[I_0, I_1), [I_1, I_2), \dots, [I_r, I_{r+1})\},$$

where  $I_s = I_0 < I_1 < I_2 < \dots < I_r < I_{r+1} = L_s$

(4)

The middle point of each interval is called a *cut*. An information feature and its corresponding cuts induced by the partition  $P_s$  is therefore uniquely represented by

$$\{(s, c_1), (s, c_2), (s, c_3), \dots, (s, c_r)\} \subset A \times \mathfrak{R},$$

where  $c_1, c_2, \dots, c_r$  are the cuts of  $[I_{i-1}, I_i]$

(5)

Using the cuts, we now construct a discrete matrix called the *discernibility table*, denoted by  $D_r^*$ . There is one column in  $D_r^*$  for each cut induced over  $D_r$ , and one row for each pair of images  $(x_i, x_j)$  where  $x_i$  and  $x_j$  have different decision classes. An entry  $v_{ij}^k$  in  $D_r^*$  is decided as follows:

$v_{ij}^k = 0$  in  $D_r^*$ , if the pair  $x_i$  and  $x_j$  of images have different decisions but the weighted value of the sub-interval  $i$  in both the images are on the same side of the cut,

$v_{ij}^k = |d_i - d_j|$  if the weight of the color sub-interval  $k$  is less than the cut in one image and greater than the cut in another image, and the images have different decision classes.

In other words, a non-zero entry corresponding to a color sub-interval in the discernibility table indicates that the color sub-interval has two different significance levels in two images of different decisions. The absolute value of the entry determines the power of the color sub-interval to distinguish between two different decision classes.

Finally, we need to find the set of *optimum color sub-intervals* from  $D_r^*$ , along with the distinguishing value of the cuts. Since, theoretically, there can be an infinite number of cuts possible, it is also required to find out a minimal set of cuts. To obtain this, we apply a modified version of the MD-Heuristic algorithm. The original algorithm as described in [7] considered all decision differences as identical. We have modified this to first consider the highest degree of difference in decision, followed by the next highest and so on, till there are no more discerning words in the set.

The steps in the modified MD-Heuristic algorithm followed by us are:

*Step 1:* Set  $T = r$ , where  $r$  is the maximum difference in decision possible. Let  $\mathcal{W}$  denote the set of optimum color sub-intervals. Set  $\mathcal{W} \leftarrow \text{NULL}$ .

*Step 2:* If there is no column containing  $T$ , then set  $T = T - 1$ ;

*Step 3:* Choose a column with the maximal number of occurrences of  $T$ 's – *this column contains a color sub-interval that is discerning the maximum number of images corresponding to the current level of discernibility*;

*Step 4:* Select the color sub-interval  $w^*$  and cut  $c^*$  corresponding to the column. Delete the column from  $D_r^*$ . Delete all the rows marked in this column by  $T$  since this discernibility is already considered. Add the color sub-interval to  $\mathcal{W}$ .

*Step 5:* If there are more rows left, then go to step 1. Else stop.

The set of optimum color sub-intervals  $\mathcal{W}$  can be used effectively to classify new set of images. In the next section we will show how a *representative sieve* is constructed from the set  $\mathcal{W}$ .

## 4 Constructing a Representative Sieve for Training Set of Images

The set of optimum color sub-intervals  $\mathcal{W}$  along with a minimal set of cuts can be used to represent training set of images of different decision classes. Using the minimal set of cuts, we now obtain a reduced decision table  $D_r$  containing optimum color sub-intervals of  $\mathcal{W}$  only. Let  $\mathcal{P}$  denote the partition

$$P = \left[ \begin{array}{l} \left\{ \left( s_1^*, c_1^{1*} \right), \left( s_1^*, c_1^{2*} \right), \dots, \left( s_1^*, c_1^{r*} \right) \right\}, \left\{ \left( s_2^*, c_2^{1*} \right), \dots, \left( s_1^*, c_2^{r*} \right) \right\}, \\ \left\{ \left( s_m^*, c_m^{1*} \right), \dots, \left( s_m^*, c_m^{r*} \right) \right\} \end{array} \right]$$

where  $c_i^{1*}, c_i^{2*}, \dots, c_i^{r*}$  are the  $i^{st}, II^{nd} \dots r^{th}$  cut for  $s_i^*$ .

(6)

For a discerning optimum color sub-interval  $s_1^*$ , we assign interval 0 to all values of  $s_1^*$  less than  $c_1^{1*}$ , interval 1 to all values lying in  $[c_1^{1*}, c_1^{2*})$ , interval 2 to all values lie in  $[c_1^{2*}, c_1^{3*})$  and so on. An analogous assignment is done for the other optimum color sub-intervals  $s_2^* \dots s_m^*$ . After this assignment, the decision table  $D_r$  is now converted to a symbolic decision table  $\mathcal{S}$  which is named as *representative sieve*.

$$\begin{aligned} \mathcal{S} &= (SW_M \cup \{d_i\}) \text{ where } 1 \leq i \leq n \text{ and} \\ SW_M &= [Sv_{ij}]_{n \times p} \text{ where } 1 \leq p \leq k, k \text{ is an integer and} \\ Sv_{ij} &= \left\{ a : a \in \text{interval no. corresponding to the minimal} \right. \\ &\quad \left. \text{cuts of } j^{th} \text{ color sub-interval of } i^{th} \text{ image} \right\} \end{aligned} \quad (7)$$

We will explain the complete approach with an example.

*Example:* We consider a set of 30 images from 3 different decision classes. In first decision class, we collect 10 images from sunset domain while in second decision class, we collect 10 images from hills or mountain, and in third decision class, we collect same no. of images from plant domain and built decision table as explained in equation 3.

The Decision table  $D_r$  is represented as

$$D_T = \left[ \begin{array}{cccccccccccccccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 & R_8 & G_1 & G_2 & G_3 & G_4 & G_5 & G_6 & G_7 & G_8 & B_1 & B_2 & B_3 & B_4 & B_5 & B_6 & B_7 & B_8 \\ 12 & 14 & 17 & 20 & 15 & 12 & 6 & 4 & 37 & 44 & 17 & 2 & 0 & 0 & 0 & 0 & 50 & 38 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\ 48 & 1 & 0 & 0 & 1 & 2 & 22 & 25 & 49 & 1 & 1 & 3 & 3 & 13 & 31 & 0 & 49 & 1 & 2 & 6 & 6 & 12 & 23 & 0 \\ \dots & \dots \\ 1 & 48 & 24 & 6 & 2 & 2 & 5 & 11 & 0 & 31 & 35 & 13 & 3 & 2 & 5 & 11 & 3 & 57 & 17 & 2 & 2 & 2 & 6 & 1 \\ \dots & \dots \\ 22 & 26 & 13 & 9 & 11 & 11 & 9 & 0 & 3 & 20 & 29 & 14 & 12 & 12 & 9 & 0 & 3 & 21 & 28 & 14 & 14 & 10 & 10 & 1 \end{array} \right] \begin{bmatrix} d_1 \\ 1 \\ 1 \\ \dots \\ 2 \\ \dots \\ 3 \end{bmatrix} \quad (8)$$

Using the approach explained earlier, the set  $\mathcal{W}$  of optimum color sub-intervals with minimal set of cuts  $c^*$  were obtained as follows:

$$W = \left\{ \begin{array}{l} s_1^* = R_8[224 - 255], s_2^* = G_1[0 - 31], s_3^* = G_3[64 - 95], s_4^* = B_5[128 - 159], \\ s_5^* = B_6[160 - 191], s_6^* = B_7[192 - 223], s_7^* = B_8[224 - 255] \end{array} \right\}$$

$$P = \left[ \left\{ [R_8, 1.5], [G_1, 11], [G_3, 26.5], [B_5, 4.5], [B_6, 14.5], [B_7, 5.5], [B_8, 4] \right\} \right] \quad (9)$$

We assign interval 0 to all values of “ $R_8$ ” less than 1.5, interval 1 to all values lying in [1.5, max. value in  $R_8$ ). As the same, we assign interval 0 to all values of “ $G_1$ ” less than 11, interval 1 to all values lying in [11, max. value in  $G_1$ ) and do it for all optimum color sub-intervals. At the end of it we thus obtain a symbolic decision table  $J$  where each sub-interval is having values in terms of interval 0, 1, 2....so on. This symbolic decision table  $J$  shows which sub-interval’s color values are contributing in classification of images. We therefore propose to use this  $J$  as a representative *sieve* to classify new set of images.

## 5 Rough Membership based Image Classification Paradigm

In this section we will present a new rough membership function to compute classification degree for new set of images. Classically, *rough membership* function quantifies the degree of relative membership of an element into a given decision class. The traditional rough-set based membership function does not work well for classification of images, since the decision table that is obtained is usually sparse. Thus we propose to use a new rough membership computation function which can take into account the contribution of all optimum color sub-intervals together. It takes into account the relative degree of membership of an image into different decision classes with respect to each color sub-interval and then provides a final categorization of new image as a function of all these memberships.

For a new image, for each optimum color sub-interval present in  $\mathcal{W}$ , the weight of color sub-interval is determined and assigned an interval to which the

weight belongs to, on the basis of minimal cuts obtained earlier is determined. The membership of new image to each decision class  $d$ , denoted by  $\mu_d(x)$ , is computed as follows. Let  $a_{s^*}(x)$  denote the interval determined by the weight of  $s^*$  for image  $x$ .

$$\begin{aligned} [s^*] &= \left\{ (x', x') \in U^2 \mid \forall s^* \in W, a_{s^*}(x) = a_{s^*}(x') \right\} \\ \mu_d^{s^*}(x) &= \frac{[s^*] \cap C}{|[s^*]|}, \text{ where } d = (1, 2, 3), \\ C &= \{x \mid x \text{ is an image having decision class } d\} \\ \forall s^* \in W, \text{ provided } [s^*] &\neq \emptyset. \mu_d^{s^*}(x) \in [0, 1] \end{aligned} \quad (10)$$

$$\mu_d(x) = \left[ \sum_j \mu_d^{s_j^*}(x) \right], \text{ where } (1 \leq j \leq p) \text{ and} \\ p = \text{no. of optimum color sub-intervals} \quad (11)$$

$$\mu_d^*(x) = \left( \mu_d(x) \Big/ \sum \mu_d(x) \right), \text{ Decision} = \left\{ d : \max(\mu_d^*(x)) \right\} \quad (12)$$

Equation (10) computes the rough membership of an image for each decision class  $d$  by taking into account the relevance of each optimum color sub-interval  $s^*$  for that decision class. Equation (11) computes the total membership value for a single decision class for new image. Equation (12) computes the final membership value for an image  $x$  to a decision class  $d$ , by normalizing it against all membership values. The decision class with the maximum weight is assigned to the new image.

## 6 Results and Discussion

In this section, we highlight the performance of the proposed approach by presenting some experimental figures. We have considered a sample set of 30 images from different decision classes like Sunset, Hills, and Plants. In table 1, we list top 5 images from each domain.

**Table 1.** List of top 5 images of each decision class from training set of 30 images

Decision Classes	Image1	Image2	Image3	Image4	Image5
Sunset					
Hills					
Plant					

To classify new set of images, we built representative sieve for a training set of 30 images and used representative sieve to classify 60 new images collected from different decision classes. Each decision class is having 20 images. To measure classification accuracy, we compare system generated final decision class of each image with user decisions for those images. In table 2, we are presenting comparison between system generated final rough memberships of top 5 images with user decisions.

**Table 2.** Comparision of system generated final membership of top 5 new images with user rated decision classes

No.	New image1	New image2	New image3	New image4	New image5
New Images					
Final Membership Value & Ratings	Class 1: 0.36 Class 2: 0.29 Class 3: 0.34 Final Class: 1	Class 1: 0.42 Class 2: 0.29 Class 3: 0.29 Final Class: 1	Class 1: 0.31 Class 2: 0.38 Class 3: 0.31 Final Class: 2	Class 1: 0.25 Class 2: 0.40 Class 3: 0.35 Final Class: 2	Class 1: 0.27 Class 2: 0.33 Class 3: 0.40 Final Class: 3
User Ratings	1	1	2	2	3

We found that 70% of new images were classified correctly. i.e. 70% of times the rough membership based decision class matches with the user rated decision class.

## 7 Conclusion

In this paper, we have proposed new rough reasoning methodologies for content based image classification. We have used the concept of discernibility to construct a representative sieve for image classification. The specific representative sieve can be act as a dedicated classifier for specific decision classes and it can be used with an intelligent agent that can constantly surf image repository for new images to classify in appropriate decision classes. We have proposed rough membership based measure for new set of images that computes contribution of each new image for each decision class and new images assigned that decision class which is having maximum rough membership value. This method is ideally suited for handling combination of multiple image features when different features may not be pointing towards the same class. Finally, we have experimented on a set of trainings images and classify new set of images with an accuracy of 70%. In future, we will extend our experiment on a large dataset to make efficient representative sieve and also classify new set of images having large number of decision classes.

## References

1. Chapelle, O., Haffner, P., Vapnik, V. N.: Support Vector Machines for Histogram-Based Image Classification. IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10(5). SEPTEMBER 1999
2. Dong, G.J., Zhang, Y.S., Fan, Y.H.: Remote Sensing Image Classification Algorithm Based on Rough Set Theory. Fuzzy Information and Engineering, Vol. (40). 846-851 (2007)
3. Komorowski, J., Polkowski, L., Andrzej, S.: Rough Sets: A Tutorial. <http://www.let.uu.nl/esslli/Courses/skowron/skowron.ps>
4. Long, F., Zhang, H., and Feng, D.D.: Fundamentals of Content-based Image Retrieval, Microsoft Research Asia publication, available online at. [http://research.microsoft.com/asia/dload\\_files/group/mcomputing/2003P/ch01\\_Long\\_v40-proof.pdf](http://research.microsoft.com/asia/dload_files/group/mcomputing/2003P/ch01_Long_v40-proof.pdf).
5. Mrowka, E., Dorado, A., Pedrycz, W., Izquierdo: Dimensionality Reduction for Content-Based Image Classification. Eighth International Conference on Information Visualisation, (IV'04) 435-438
6. Mohabey, A., Ray, A. K.: Rough set theory based segmentation of color images. Fuzzy Information Processing Society, 2000. NAFIPS. 19th International Conference of the North American, 338 – 342(2000)
7. Park, S. B., Lee, J. W., Kim, S.K.: Content-based image classification using a neural network. Pattern Recognition Letters, Vol. 25(3). (2004) 287-300
8. Pawlak, Z.: Rough Sets. Int. Journal of Computer and Information Sciences, Vol. 11(5). 341-356(1982)
9. Sergyán, S.: Color Content-based Image Classification. 5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics, January 25-26, 2007

10. Shang, C., Shen, Q.: Aiding Neural Network Based Image Classification with Fuzzy-Rough Feature Selection. 2008 IEEE International Conference on Fuzzy Systems (FUZZ 2008)
11. Shih, J., L., Chen, L.W.: A Context-Based Approach for Color Image Retrieval, International Journal of Pattern Recognition and Artificial Intelligence, Vol. 16 (2). 239-255. (2002)
12. Singh, S., Dey, L.: A new Customized Document Categorization scheme using Rough Membership. International Journal of Applied Soft-Computing, Vol. 5 (4), 373-390(Jul 2005)
13. Vailaya, A., Figueiredo, M.A.T., Jain, A.K., Hong-Jiang Zhang: Image classification for content-based indexing. Image Processing, IEEE Transactions on Vol. 10 (1). 117 – 130(Jan 2001)
14. Wang, S.L., Liew, A.W.C.: Information-Based Color Feature Representation for Image Classification. IEEE International Conference on Image Processing (ICIP 2007), Vol. 6. 353 - 356(Sept.2007)
15. Yun. J., Zhanhuai, L., Yong W., Longbo Z.: A Better Classifier Based on Rough Set and Neural Network for Medical Images. Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06) 853-857(2006)

# A Robust Object Tracking Method for Noisy Video using Rough Entropy in Wavelet Domain

Anand Singh Jalal and Uma Shanker Tiwary

Indian Institute of Information Technology, Allahabad, India  
(asjalal@iiita.ac.in, ust@iiita.ac.in)

**Abstract.** In this paper we have proposed a robust object tracking method using rough entropy and flux in wavelet domain. The tracking framework necessitates robust and efficient but accurate methods for segmentation and matching. The object is represented in wavelet domain features to minimize the effect of frame to frame variations and noise. The concept of maximizing rough entropy in wavelet domain helps in finding out the threshold value to make a distinction between the object and the background pixels in a vague situation. The search for the candidate subframe is made fast by using the motion prediction algorithm. A measure based on flux in wavelet domain combined with the number of pixels in the object has been developed. The proposed tracking algorithm yields better results even in noisy video as shown in the experiments. The results show that the wavelet domain segmentation and tracking improves the localization error approximately by 5-7%.

## 1 Introduction

Tracking an object in a complex environment is a challenging task [1]. The problem become hard when applied to real life situations such as sport video analysis to extract highlights, surveillance system to know the suspicious activity, traffic monitoring, and human computer interface to assist visually challenged people [2, 3]. Most of the algorithms for tracking envisage a controlled laboratory environment. The algorithms fail completely in the presence of noise in the video or the movement is complex or the object shape is irregular, etc. In this paper, an attempt has been made to develop a robust method to yield the result even in noisy conditions. The task of object tracking in a video comprises of three sub tasks: Segmentation, Searching possible location of candidate object in each frame and Localization through a similarity measure.

The aim of segmentation algorithm is to discriminate the object from the background. A mean-shift approach to compute clusters using both spatial and color features was proposed by Comaniciu and Meer [4] for image segmentation in tracking. Shi and Malik [5] gave a method known as normalized graph cut to deal with the over-segmentation problem of mean-shift method. Pal et al. [6] has used the rough entropy concept to get the threshold value to segment the object and the background. One of the brute force methods may be to exploit the object detection technique on each possible subframe of every frame. But this puts an overhead of exhaustive search. In general it is an observed fact that compared with the target object in the previous frame; the candidate object in the next frame usually does not

change abruptly. In [7] kalman filter is used to estimate the state of a linear system. However, heavy computational burden is one of the major problems with kalman filter method. Features play an important role in the performance of a tracking algorithm. Selecting features that are able to discriminate between multiple objects and the background and are also good for tracking is a hard problem. The mean shift algorithm based on color has recently been proposed as an efficient algorithm for object tracking [8, 9, 10]. In their work they use Bhattacharyya coefficient as a measure of comparability between target information in previous frame and candidate information in the next frame and optimize the search using mean shift technique. One of the problems using color as feature is that it will create confusion when the background varies. Also it is sensitive towards the illumination changes [11]. In recent years the wavelet feature based techniques have gained popularity in object tracking [12, 13]. In [12] the author uses highest energy coefficients of Gabor wavelet transform to model the object in the current frame and 2D mesh structure around the feature points to achieve global placement of the feature point. The 2D golden section algorithm is used to find the object in the next frame. Most of these algorithms suffer with a localization error that will rise as the background changes.

In this paper we have proposed a robust object tracking technique using rough set and Daubechies wavelet. In the proposed algorithm the attempt has been made to minimize the localization error due to the background pixels. The location information of previous three frames is used to compute the current position [14, 15]. The proposed tracking algorithm uses the concept of rough set for obtaining threshold for object-background classification. Rough entropy provides a better framework to handle uncertainty in the object and background. The discrete wavelet transform (DWT) is the tool of choice when we require to view and process a digital image at multiple resolutions [15]. The DWT also gives insight into frequency and spatial characteristics of image. In this paper Daubechies wavelet coefficients at first level are used to represent the object. Tracking at this low resolution is computationally less expensive and less sensitive to noise. The flux in wavelet domain has been considered as a measure of comparability between the target object and the candidate object. The proposed method is robust as it is also able to track the object in noisy video sequences.

The rest of the paper is organized as follows: Section 2 describes some background concepts, Section 3 describes the proposed algorithm. The experimental results and discussion are given in section 4 and finally the conclusions are given in section 5.

## 2 Background Concepts

In this section we present some background concepts used in the proposed algorithm.

## 2.1 Rough Entropy

Pal et al. in [6] has given an algorithm on gray value of the image using rough set to find out a threshold to segment the object and background. In their work they have used the concept of rough entropy with respect to a threshold value defined as

$$RE_{Th} = -\frac{e}{2}[(ROF_{Th}) \log_e(ROF_{Th}) + (RBF_{Th}) \log_e(RBF_{Th})] \quad (1)$$

Where  $ROF_{Th}$  and  $RBF_{Th}$  are the probability of roughness of the object and background respectively. The value of the threshold is calculated by finding the argument of maximum entropy as

$$Th^* = \arg \max_{Th} RE_{Th} \quad (2)$$

The probability of roughness is defined as the complement of accuracy in rough set, where accuracy is represented as ratio of lower approximation and upper approximation.

$$ROF_{Th} = 1 - \text{accuracy}( ) = 1 - \frac{|OF_{Th}|}{|OF_{Th}|} \quad (3)$$

These lower and upper approximations can be found out for the whole image of interest. It can also be found out as in [6] by dividing the image into granules of proper size and aggregating the lower and upper approximation of each granule as defined below:

The lower approximation ( $OF_{Th}$ ) and upper approximation ( $\overline{OF}_{Th}$ ) of the object is given as

$$(OF_{Th}) = \bigcup_i gf_i, \quad j=1, \dots, rc \quad \overline{OF}_{Th} = \bigcup_i gf_i \quad j=1, \dots, rc$$

Such that  $P_j > Th$ , Where  $P_j$  is the probability that the selected gray value is greater than threshold and hence part of the object.

The lower approximation ( $BF_{Th}$ ) and upper approximation ( $\overline{BF}_{Th}$ ) of the background is given as

$$(BF_{Th}) = \bigcup_i gf_i, \quad j=1, \dots, rc, \quad \overline{BF}_{Th} = \bigcup_i gf_i \quad j=1, \dots, rc$$

Such that  $P_j \leq Th$ , Where  $P_j$  is the probability that the selected gray value is equal or smaller than threshold and hence part of the background.

However in actual calculation the threshold value is unknown and is calculated iteratively by exploring the possibility of each value between minimum and maximum feature value.

## 2.2 Wavelet Representation

The discrete wavelet transform (DWT) provides an efficient framework for representation and storage of images at multiple levels [15]. One of the features of DWT is that the spatial information is retained even after decomposition of an image into four different frequency bands. In this decomposition, the high frequency sub images (horizontal coefficient, vertical coefficient and diagonal coefficient) contain the detailed information. In a moving object tracking framework the global information i.e. low frequency part is sufficient to locate the object. Some of the abrupt changes in object or background conditions get smoothed in the approximation coefficient sub image, which gives robustness property to the algorithm. This may include a resolution loss but that is not very critical in object tracking.

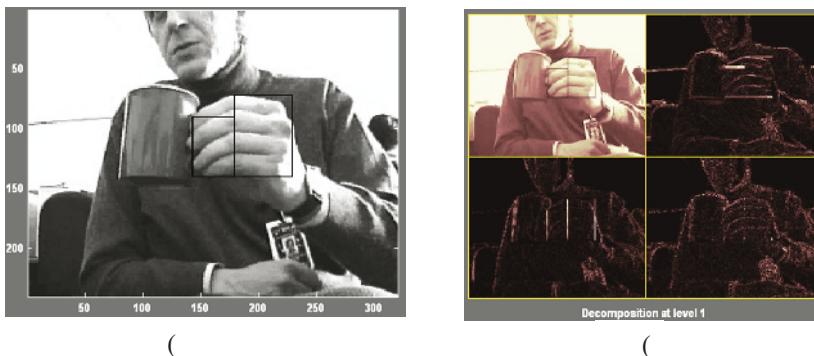


Fig. 1: DWT a) Original image      b) The result of one scale DWT

One way to resolve the noise problem is to denoise each frame. This will be computationally expensive. The wavelet transform provides a robust framework that is capable to handle noise. Also as the size of wavelet coefficient is one fourth of the original image, the search area becomes small.

## 2.3 Motion Prediction

The tracking algorithm assumes that in a few consecutive frames the trajectory of object does not change abruptly. We compute the current position based on the position, velocity and acceleration information of the previous three frames as

$$X_i = X_{i-1} + V_{i-1} + \frac{1}{2} A_{i-1} \quad (4)$$

$$Y_i = Y_{i-1} + V_{i-1} + \frac{1}{2} A_{i-1} \quad (5)$$

Where  $X_i$ ,  $Y_i$  represent the position in (x,y) coordinate and  $V_i$  ( $A_i$ ) and  $\frac{1}{2} A_i$  ( $\frac{1}{2} A_i$ ) represent the velocity and acceleration respectively along x(y) direction and can be computed as

$$\begin{aligned} X_i &= X_i - X_{i-1} \\ {}^2X_i &= X_i - 2X_{i-1} + X_{i-2} \text{ for all } i > 2 \end{aligned}$$

This prediction method has been used to reduce the search space and works fairly accurately even for complex motion.

### 3 The Proposed Tracking Algorithm

The proposed tracking algorithm uses the concept of rough set for obtaining threshold for object-background segmentation. Once the frames of video are extracted, a target object with some portion of the background is chosen. We represent this object using a primitive shape e.g. a rectangle. One can select more than one rectangles or other shapes to represent the complex shaped object accurately. Then, the feature vector of this representation is calculated i.e. first level Daubechies wavelet approximation coefficients of object points in the selected area. We also compute a global parameter of the object called flux, by aggregating these values of the wavelet coefficients. The next task is to locate the candidate object in the next frame. The position of the object in the next frame is predicted by equation (4) and equation (5). Candidate subframes of the object are generated by shifting the predicted position by  $\pm p$  points. By comparing the target and candidate feature vectors, a pair with maximal similarity can be found. The same process is repeated for all (or alternate) frames of the video.

In the proposed algorithm to compute the flux we are using low frequency sub-image (approximate coefficient) wavelet coefficients for target and candidate representation. In this computation of flux we are considering only pixels which belong to the object. This can be done by obtaining threshold for object-background segmentation using the rough entropy method described in section 2.1.

The proposed algorithm is as follows:

Step 1: Make a large square bounding box to cover the object and some portion of the background. Compute the Daubechies wavelet coefficient of the pixels in the bounding box. Use rough set segmentation algorithm to find threshold ( $Th$ ) between object and background in the wavelet domain.

Step 2: Let the centroid of bounding box covering the object ( $T\_Obj$ ) in first frame is at  $(C_x, C_y)$ . Compute the flux of Daubechies wavelet coefficients of the bounding box, say  $F$ , and the number of feature values  $N$  that are used in computation of flux i.e. those feature values which are greater than or equal to  $Th$ :

$$F = \sum_{(i,j)} w_{i,j} \quad i, j, \text{ s.t. } w_{i,j} > Th$$

and  $\mathbf{N} = |\mathbf{F}|$

Where  $w_{i,j}$  is 2D Daubechies wavelet approximation coefficients at  $(i,j)^{\text{th}}$  point.

```

Step 3: For Frame_No=2 to End_of_Frame_Sequence do
    Compute the Daubechies wavelet coefficient of the
    frame,
    if Frame_No < 4
        search_length(SL) = 2p
    else
        search_length(SL) = p
        With the help of equation(4) and
        equation(5) predict the centroid ( $C_x$ ,  $C_y$ ) of
        the current frame.
    endif

    for i= -SL to +SL do
        for j= -SL to +SL do
             $C_{\text{newx}} = C_x + i$ ;  $C_{\text{newy}} = C_y + j$ ;
            Make a bounding box with centroid ( $C_{\text{newx}}$ ,
             $C_{\text{newy}}$ ), which is the subframe of the candidate
            object  $C_{\text{Obj}}$ . Compute the difference of flux
            of  $C_{\text{Obj}}$  and  $T_{\text{Obj}}$ , say  $dF_{i,j}$ . Also compute
            the difference of N, say  $dN_{i,j}$ 
        end
    end

    Find the index of the minimum  $dF_{i,j}$ , say  $(m_1, n_1)$ 
    if for  $\{dF_{i,j}\}$  the  $\{dN_{i,j}\}$  is less than 5
    percentage of N then
         $(m, n) = (m_1, n_1)$ 
    else
        Find the index of minimum  $dN_{i,j}$ , say  $(m_2, n_2)$ 
         $(m, n) = (m_2, n_2)$ 
    endif
     $C_x = C_x + m$ ;  $C_y = C_y + n$ ;
end.

```

## 4 Experimental Results and Discussion

To show the performance of the proposed object tracking method, we have used a video consisting of 99 frames (each of size 240 x 320) to track hand movement [16].

To compare our proposed algorithm, we have implemented another algorithm, which is similar to our proposed method but without taking wavelet transform.

#### 4.1 Localization Error

To measure the performance, the actual position of the object in each frame is marked manually and the euclidian distance between the centroid positions in tracked object frame and the corresponding hand marked frames is calculated. We are calling this as localization error. The localization error of the proposed wavelet method and without wavelet method is shown in Figure 2(b). It is worth mentioning that if the localization error exceeds the window size ( $2p+1$ ), the search window will go out of the object subframe in subsequent frames and the tracking halts. Thus to keep the localization error within bound another constraint on the flux based similarity has been added, i.e. if  $N$  varies more than 5%, then realign object with minimum variation in  $N$ . Figure 2(b) shows that the wavelet domain segmentation and tracking improves the localization error approximately by 5-7%.

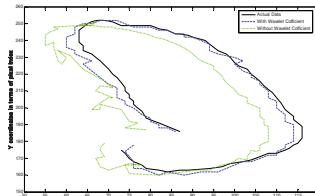


Fig. 2 (a). Trajectory of object movement in each frame

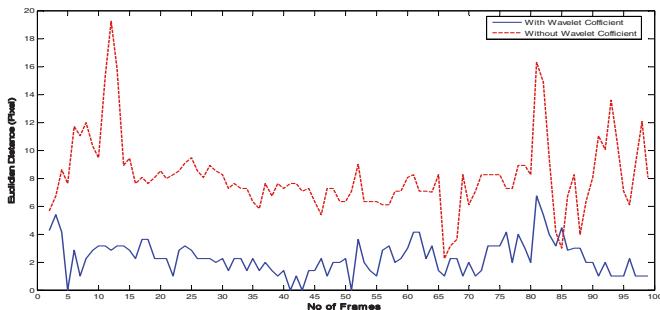


Fig. 2 (b). The error of object location in each frame using the euclidian distance measure (in number of pixels)



Frame 12

Frame 43

Fig. 2 (c). Results of tracking hand in 12<sup>th</sup> and 43<sup>rd</sup> frame respectively, the dark bounding box shows the results with wavelet features and light bounding box shows the results without wavelet features.

From the figures 2(a), 2(b) and 2(c), it is clearly shown that the wavelet based method is better. Figure 2(b) shows the localization error, which can be calculated as  $\text{SQRT}((X-X_i)^2 + (Y-Y_i)^2)$ , where  $(X,Y)$  is the centroid of actual object and  $(X_i,Y_i)$  is the centroid of tracked object.

#### 4.2 Effect of Noise

The tracking algorithm on gray scale images are very sensitive to noise and changes in the parameters such as illumination, texture, etc. The wavelet domain thresholding and feature calculation reduces some of these errors, as wavelet domain thresholding is widely used for denoising [17]. Figure 3(a), 3(b) and 3(c) show the tracking error by including white Gaussian noise in the video sequences. The same measures are used for comparison as used in Figure 2(b). From the figures 3(a), 3(b) and 3(c) it is clear that wavelet based version of the proposed algorithm gives better results even in the presence of noise.

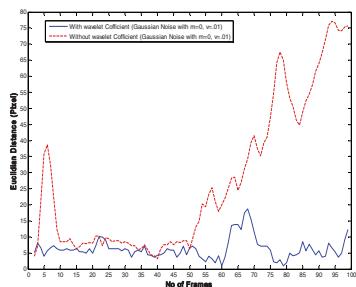


Fig. 3(a)

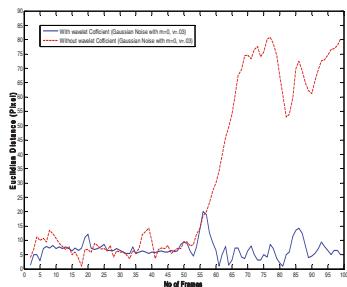


Fig. 3(b)

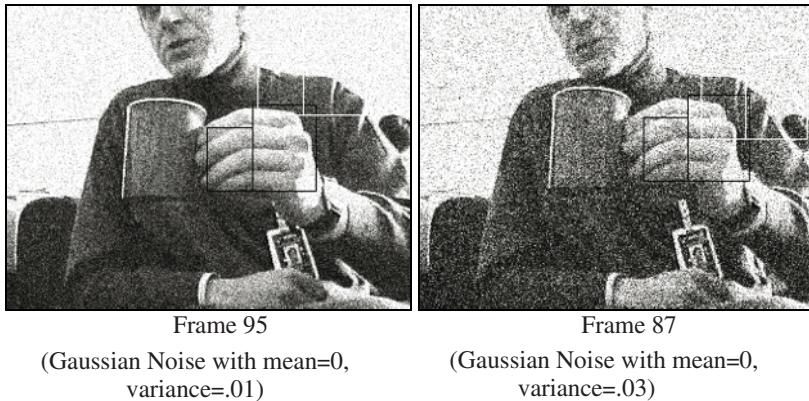


Fig. 3c

Fig. 3(a) and Fig. 3(b) shows the error of object location using the euclidian distance (in number of pixels) in each frame having a) Gaussian Noise with mean=0, variance=.01, b) Gaussian Noise with mean=0, variance=.03

Fig. 3(c) shows Results of tracking hand using noisy video in 95<sup>th</sup> and 87<sup>th</sup> frames respectively, the dark bounding box shows the results with wavelet features and light bounding box shows the results without wavelet features

## 5 Conclusion

Discrete wavelet transform (DWT) do not lose the spatial information even after decomposition of an image into four different frequency bands. In this paper a new robust algorithm is proposed for object tracking using wavelet based features. To improve the results the rough entropy measure has been used to discriminate the background pixels from the object pixels. The position, velocity and acceleration information from previous three frames are used to predict the location of object in the current frame, which gives fairly accurately predictions even for complex motion, so that we could keep a search window of  $\pm 3$ . Also we are keeping record of no of feature values used in object model. The result shows that wavelet based approach gives better results even in the presence of Gaussian (upto variance of .03) noise. The future work may include the use of complex wavelet to make this algorithm shift, scale and orientation invariant. Also work can be done to get edge information which can be included as additional features for tracking.

## References

1. Sonka M., Hlavac V., Boyle R.: *Image Processing, Analysis and Machine Vision*. Thomson Asia Pvt. Ltd., Singapore (2001)
2. Hu W., Tan T., Wang L., Maybank S.: A survey on visual surveillance of object motion and behaviours. *IEEE Transactions on Systems, Man and Cybernetics*, vol. 34, No. 3 (2004)
3. Yilmaz A., Javed O., Shah M.: Object Tracking: A Survey. *ACM Journal of Computing Surveys* 38(4) (2006)
4. Comaniciu D., Meer P.: Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Patt. Anal. Mach. Intell.*, 603–619 (2002)
5. Shi J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. Patt. Anal. Mach. Intell.* 22, 8, 888–905 (2000)
6. Sankar K. Pal, Uma Shankar B., Pabitra Mitra: Granular computing, rough entropy and object extraction. *Pattern Recognition Letters*, 26(16): 2509–2517 (2005)
7. Broda T., Chellappa R.: Estimation of object motion parameters from noisy images. *IEEE Trans. Patt. Anal. Mach. Intell.* 8, 1, 90–99 (1986)
8. Comaniciu D., Ramesh V., Meer P.: Real-time Tracking of non-rigid objects using Mean Shift. *IEEE Conference on Computer Vision and Pattern Recognition*, South Carolina, pp. 142–149 (2000)
9. Comaniciu D., Ramesh V.: Mean Shift and optimal prediction for efficient object tracking. *Proceedings of the IEEE Int'l Conference on Image Processing (ICIP)*, pp.70-73 (2000)
10. Comaniciu D., Ramesh V., Meer P.: Kernel-based object tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25 (5), pp. 564–575 (2003)
11. Khansari M., Rabiee H. R., Asadi M., Ghanbari M.: Object tracking in crowded video scenes based on the Undecimated wavelet features and texture analysis. *EURASIP Journal on Advances in Signal Processing*, Article ID 243534 (2008)
12. He C., Zheng Y. F., Ahalt S. C.: Object tracking using the Gabor wavelet transform and the golden section algorithm. *IEEE Transactions on Multimedia*, vol. 4, no. 4, pp. 528–538 (2002)
13. Feris R. S., Krueger V., Cesar Jr. R. M.: A wavelet subspace method for real-time face tracking. *Real-Time Imaging*, vol. 10, no. 6, pp. 339–350 (2004)
14. Ashish Khare, Uma Shanker Tiwary: Daubechies complex wavelet transform based moving object tracking. *Proceedings of the IEEE Symposium on Computational Intelligence in Image Processing*, USA, pp. 36-40, 1–5 April (2007)
15. Yiwei Wang, John F. Doherty, Robert E. Van Duck: Moving object tracking in video. *Proceedings of 29<sup>th</sup> IEEE Int'l Conference on Applied Imagery Pattern Recognition Workshop*, pp. 95–101 (2000)
16. [www.gergltd.com/cse486/project5/images/redcup.avi](http://www.gergltd.com/cse486/project5/images/redcup.avi)
17. Khare A., Tiwary, U.: Soft-Thresholding for Denoising of Medical Images: A multiresolution approach. *International Journal of Wavelets, Multiresolution and Information Processing (IJWMIP)*, 3, 477–496 (2005)

# Human Computer Interaction

# Extraction of Rhythmic Information from Non-invasively Recorded EEG Signal Using IEEE Standard 1057 Algorithm

Manoj Kumar Mukul and Fumitoshi Matsuno

Department of Mechanical Engineering and Intelligent Systems, The University of  
Electro-Communications, Tokyo, Japan  
fmkm, matsunog@hi.mce.uec.ac.jp

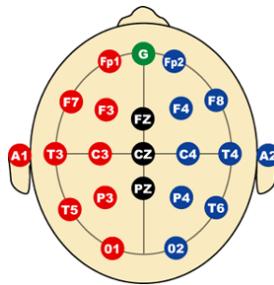
**Abstract.** This paper describes extraction of rhythmic activity from one or two channel non-invasively recorded signal using IEEE standard 1057 algorithm. However BSS technique such as ICA is generally applied to Multichannel recordings and the analysis of single channel recordings with BSS technique is not usually performed. However there are instances where just one or two recording channel is either available or desired the difficulty of isolating signals of interest is greatly increased. It would be particularly useful to be able to automatically isolate visualize and track multiple neurophysiologically meaningful sources under laying the ongoing single or two channel EEG recording. We introduced a method whereby it is possible to break down single or two channels recordings of EEG brain signals into their underlying components, irrespective of the components origin, method relies on a combination of a nonlinear dynamical system framework using standard implementation of IEEE standard algorithm.

## 1 Introduction

In the analysis of noninvasively recorded brain signals it is required to extract neurophysiologic ally meaningful information either for clinical reasons or communications. Similarly EEG can be used as in the field of Brain- Computer Interfacing (BCI) [1] where brain signals are interpreted to provide a means of communication. A powerful technique in the decomposition of multi-channel EEG brain signals is the technique of Blind Source Separation (BSS), [2] in particular recent efforts in Independent Component Analysis (ICA) to this end. However, BSS techniques such as ICA are generally applied to multi-channel recordings and the analysis of single channel recordings with BSS techniques is not usually performed. However, there are instances where just one or two recording channel is either available or desired; the difficulty of isolating signals of interest is greatly increased. In general, rhythmic activity in the EEG is of interest, ( $\alpha$ ,  $\beta$ ,  $\delta$  and  $\gamma$  band activities, or rhythmic seizure activity). It would be particularly useful to be able to automatically isolate, visualize and track multiple neurophysiologic ally meaningful sources underlying the ongoing single or

two channels EEG recording. We introduced a method whereby it is possible to break down single or two channel recordings of EEG brain signals into their underlying components, irrespective of the components' origin. The method relies on a combination of a nonlinear dynamical systems framework using standard implementation of IEEE standard algorithm 1057 [3, 4]. The brain state of the individual may make certain frequencies more dominant. The best-known and most extensively studied rhythm of the human brain is the normal alpha rhythm (8-13[Hz]). It can be usually observed better in the posterior and occipital regions with typical amplitude about 50  $\mu$ V (P-P). Alpha activity is induced by closing the eyes and by relaxation, and abolished by eye opening or alerting by any mechanism (thinking, calculating). Useful information can also be extracted from particular brain configurations that can be interpreted in terms of brain states. This paper describes the extraction of  $\alpha$  and  $\beta$  rhythmic activity from non-invasively recorded signal in time domain. Extraction of rhythmic activity from non-invasively recorded signal is a kind of filtering. Also this paper compares the effect of FIR filtering as well as Non-linear Square filtering on bipolar recorded EEG signal.

## 2 Data Acquisition



**Fig. 1.** 10-20 Electrode Placement

The Graz group [5] has developed a BCI which uses  $\mu$  (8-12 Hz) and central  $\beta$  (18-25 Hz) EEG rhythms recorded over the motor cortex. Several factors have suggested that  $\mu$  and/or  $\beta$  rhythms may be good signal features for EEG-based communication. The data was recorded from two subjects (S1 and S2) over two sessions, in a timed experimental recording procedure. Each trial was 20[s] length. The subject was verbally asked to perform the mental task related to imagination of movement of left, right hand and straight movement. For each subject a total of 6 trials were recorded (3 trials of each type of movement imagery). The recording was made using a g.tec amplifier (<http://www.gtec.at/>) and Ag/AgCl electrodes. All signals were sampled at 256[Hz] and filtered between 0.5 and 30[Hz]. Two bipolar EEG channels were measured using two electrodes positioned 2.5 cm posterior ("−") and anterior ("+") to position C3 and C4 according to the international standard (10/20 system) electrode positioning

nomenclature which has been shown in Fig. 1. In bipolar recording the recorded voltage is the voltage difference between the anterior and posterior electrode at each recording site. The recorded data consists of sample points having the p number of real valued frequency components contaminated by an additive measurement noise. Each frequency component can be modeled as

$$s_i[n] = A_i \cos(\omega_i n) + B_i \sin(\omega_i n) + C_i = \alpha_i \sin(\omega_i n + \Phi_i) + C_i \quad (1)$$

The constant  $A_i$ ,  $B_i$ , and  $C_i$  are all assumed to be unknown. The constant angular frequency are also considered known / unknown parameters. Where  $A_i = \alpha_i \sin(\Phi_i)$  and  $B_i = \alpha_i \cos(\Phi_i)$  the measured signal  $x[n]$  is sum of the p frequency components and an additional noise term  $w[n]$ , that is

$$x[n] = \sum_{i=1}^p S_i[n] + w[n] \quad (2)$$

The noise is assumed to be zero mean white Gaussian noise with variance  $\sigma^2$ . Accordingly the data Model is given by

$$x = H\theta + w \quad (3)$$

$H$  is  $(N \times 2p + 1)$  matrix

$$H = \begin{bmatrix} 1 & \cos\omega_1 & \sin\omega_1 & \cdots & \cos\omega_p & \sin\omega_p \\ 1 & \cos 2\omega_1 & \sin 2\omega_1 & \cdots & \cos 2\omega_p & \sin 2\omega_p \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & \cos N\omega_1 & \sin N\omega_1 & \cdots & \cos N\omega_p & \sin N\omega_p \end{bmatrix} \quad (4)$$

The parameter has to be estimated as gathered in  $\theta$  as

$$\theta = [C \ A_1 \ B_1 \ A_2 \ B_2 \ \cdots \ A_p \ B_p]^T, (C = \sum_{i=1}^p C_i) \quad (5)$$

$C$  has been the sum of individual dc components

Known and unknown frequency parameters has to be estimated

$$\omega = [\omega_1 \ \omega_2 \ \cdots \ \omega_p]^T \quad (6)$$

### 3 Extractions of Rhythmic Components

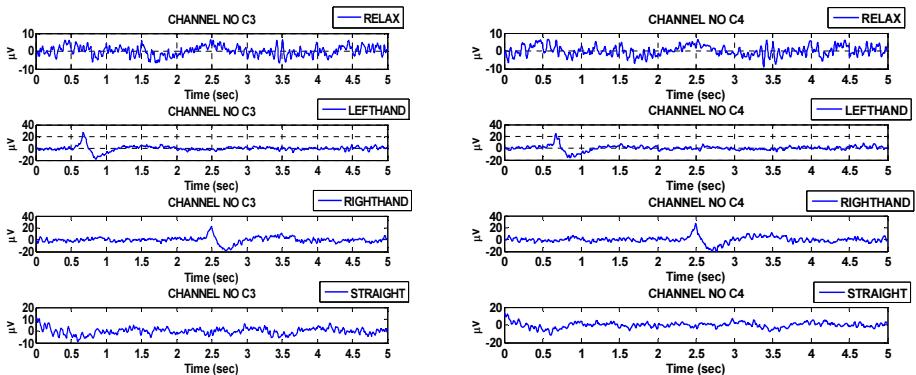
EEG signal is non-stationary random signal. The above mentioned approach is not directly applied to non-stationary random process. In order to reconstruct the signal from the pseudo randomly taken sample values by applying the un-orthogonal transform equation 7. The number  $N$  of the signal samples taken

has to satisfy condition which is given as  $N \geq 2p$ . We are going to extract the components having the frequency range which covers the whole  $\alpha$  and  $\beta$  rhythm of EEG signal. For that we have segmented the whole recorded data having the segment of 1[sec]. We have the sampling frequency is 256[Hz]. We have taken the 256 samples for constructing the sinusoidal signal which corresponds to rhythmic range (8-30[Hz]). In each segment we have obtained the magnitude and phase of same frequency components and stacked in succession. In order to remove the abrupt discontinuity at start of each segment we have taken the concept the overlapping and averaging technique. Number of samples which is to be shift falls within the frame length. It cannot be greater than number of samples in frame. Mathematically the obtained value of  $X(\lambda, f)$  for each segment is given by the equation

$$X(\lambda, f) = (H^T H)^{-1} H^T x \quad (7)$$

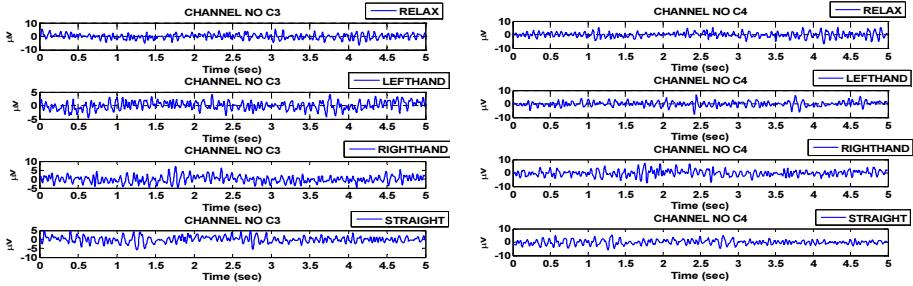
## 4 Result

In order to extract the rhythmic components related to  $\alpha$  and  $\beta$  rhythm from recorded time domain signal, we have recorded the data from two subjects related to mental tasks Relax, Lefthand, Righthand and Straight movement. The plotting of the recorded 5[sec] signal from the motor cortex area having the electrode placement C3 and C4 according to 10-20 electrode placement is shown in the Fig. 2.



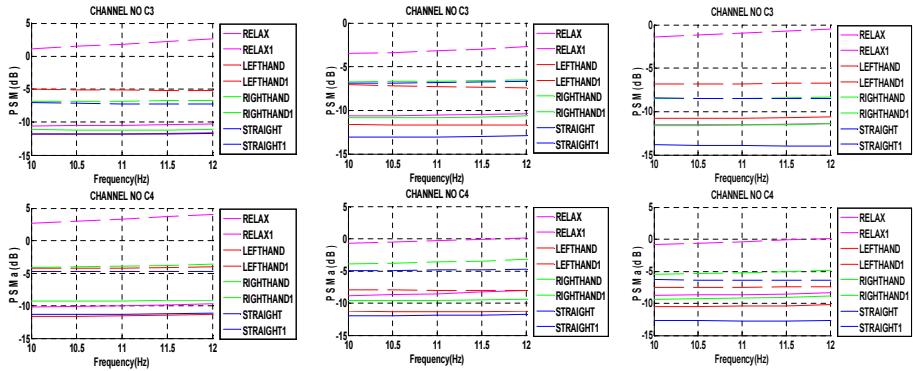
**Fig. 2.** Plotting of 5S Recorded Signal

The signal obtained after passing through the algorithm is shown in the Fig. 3. The power spectral density in the Motor rhythm (10-12[Hz]) revels clear cut differences in the mental task performed by the subject. In the case of subject 2 there is also coming the contralateral effect i.e. the right side of the brain is going to control left side of the body and Left side of the brain is going to control right side of the body.



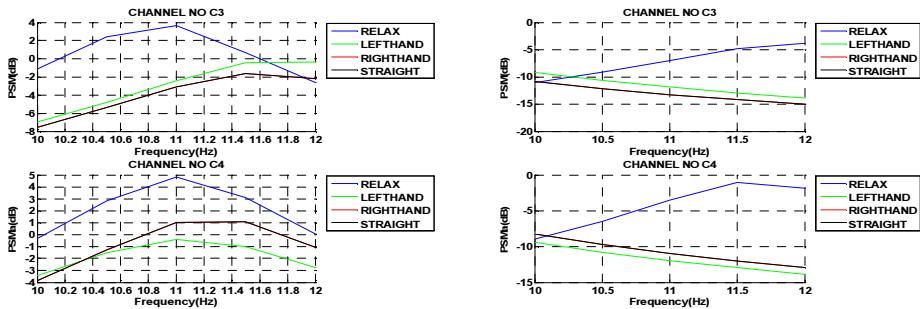
**Fig. 3.** Plotting of Extracted signal having the components 8-30[Hz]

Fig. 4 shows the comparative PSD curve for subject 1 having the Legend Relax, Lefthand, Righthand and Straight as well as for subject 2 having the legend Relax1, Lefthand1, Righthand1, and Straight1 for mental task. In among the three trials the trail No. 2 and 3 show the similar kind of characteristics for both subjects except in power level in Motor Rhythm. We conclude that above proposed algorithm most appropriate for extraction of rhythmic information from single channel recording.



**Fig. 4.** PSD for Trial 1, 2 and 3 in Motor Rhythum

There are clear cut differences among the mental task performed by the subject after extraction of signal component having the frequency range 8-30[Hz] using above mentioned algorithm for non-stationary signal rather than FIR filtering. By comparing the Fig. 5 for both subject similar kinds of characteristics have been obtained but the power level for Straight movement as well as right hand movements have completely overlapped. For classifier design using the characteristics feature as Motor rhythm using above mentioned algorithm as the preprocessing technique will greatly simplified the design aspect of classifier by taking the benefit of constant power level differences in whole frequency rhythm.



**Fig. 5.** PSD of subject S1 and S2 after FIR Filtering

## 5 Conclusion and Future work

This paper presents noise cancellation i.e. removal of noise signal which can be either EMG, ECG or a combination of these two artifacts from the corrupted EEG signal and also signal enhancement both using the information rhythm which contains the maximum information related to mental task. For this purpose, we have implemented the IEEE-STD 1057 four parameter sin wave fit algorithm, which is the most recent and sophisticated real time signal construction algorithm. Thus there is a need of faster way of calculation which reduces the computational time. Finally we are planning to design the classifier using the characteristics features as power spectral density for Motor rhythm and development of translation algorithm which in turns converts the classified features in control command for robotic movement using single or two channel data in real time operation.

## References

1. Wolpaw, J.R., McFarland, D.J., Vaughan, T.M.: Brain-computer interface research at the wads worth center. *IEEE Transaction on Rehabilitation Engineering* (2000)
2. Cichocki, A.: Blind signal processing methods for analyzing multichannel brain signals. *Journal of Bioelectromagnetism* (2004)
3. Handel, P.: Properties of the ieee-std-1057 four-parameter sine wave fit algorithm. *IEEE Transaction on Instrumentation and Measurement* (2000)
4. Anderson, T.: Multiple-tone estimation by ieee standard 1057 and the expectation-maximization algorithm. *IEEE Transaction on Instrumentation and Measurement* (2005)
5. Coyle, D.: A time-frequency approach to feature extraction for a brain-computer interface with a comparative analysis of performance measures. *Euraship Journal on Applied Signal Processing* (2005)

# Registration of Multimodality Medical Imaging of Brain using Particle Swarm Optimization

Mahua Bhattacharya<sup>1,\*</sup> and Arpita Das<sup>2</sup>

<sup>1</sup>Indian Institute of Information Technology & Management, Gwalior Morena Link Road, Gwalior-474003, India

\* Corresponding author e-mail: [mb@iiitm.ac.in](mailto:mb@iiitm.ac.in)

<sup>2</sup>Institute of Radio Physics & Electronics, University of Calcutta  
92, A.P.C. Road, Kolkata-700009

<sup>2</sup>e-mail: [dasarpita\\_rpe@yahoo.co.in](mailto:dasarpita_rpe@yahoo.co.in)

**Abstract:** In present work we have introduced nonlinear affine registration method to incorporate the anatomic body deformation. The present technique has been developed for registration of section of human brain using CT and MR modalities. Present study related to image registration of different modality imaging is based on 2-D/2-D affine registration technique. Automatic registration has been achieved by maximization of a similarity measure and which is the correlation function of two images. The proposed method has been implemented by choosing a realistic, practical transformation and optimization techniques. Since similarity metric is a non-convex function and contains many local optima, choice of search strategy for optimization is important in registration problem. There exist many optimization schemes, most of which are local and require a starting point. Presently, we have implemented multiresolution based particle swarm optimization technique to overcome this problem.

## 1 Introduction

Medical imaging provides insights into the size, shape and spatial relationships among anatomical structures. In radiotherapy planning, dose calculation is based on the computed tomography (CT) data, while tumor outlining is often better performed in the corresponding magnetic resonance imaging (MRI). These images are used in a complimentary manner to gain additional insights into the phenomenon. The different modalities must be appropriately combined or fused to extract more useful information for diagnosis from the fused data [2, 16]. Before images can be fused, they must be geometrically aligned. This alignment process is known as registration [1, 2]. We have already done experiment on

image registration using MR (  $T_1$  and  $T_2$  weighted both ) and CT imaging modalities of ventricular region of human brain for patients having Alzheimer's diseases and other neurodegenerative diseases using shape theoretic approach [2, 17]. The control points on the concavities present in the contours are chosen to re-project ROI from the respective modalities in a reference frame [2]. In different works of image registration gray value correlation has been applied and all these information are further used in the images to determine the best match. The resulting matching is fully automatic and assumed to be maximal if images are geometrically aligned [2–5]. Proposed algorithm works using 2D/2D affine registration of multimodality (MR & CT) brain imaging. The focus of the current paper is the search strategy (optimization) for maximization of the similarity metric. The similarity metric is generally not a smooth function and contains many local optima [6] hence the choice of optimization routine plays a key role in the process of registration. In present approach we have introduced Particle Swarm Optimization (PSO) for search strategy for maximization of similarity metric. Particle swarm optimization was discovered through the simulation of a simplified social model, more specifically the collective behaviors of simple individuals interacting with their environment and with each other. In theory at least individual members of the flock can profit from the discoveries and previous experience of all other members of the flock during the search for food. This hypothesis was fundamental of the development of PSO [7–9]. Optimization schemes are performed in a multiresolution manner to decrease the sensitivity of the method to local maxima during the process of registration. The initial orientation of the images to be registered can be obtained from the reduced resolution [10, 11].

## 2 Methodology for Image Registration

In present methodology for image registration we have described *particle swarm optimization* PSO based approach in multiresolution domain for optimization of the similarity metric. Proposed methodology determines correlation based affine transformation of floating images related to similarity metric and maximizes it to achieve appropriate registration process implemented on CT and MR images of section of human brain.

### 2.1. Transformation

The transformation technique applied to register the images can be categorized according to the degrees of freedom. Although elastic transformations are more realistic (as most body tissues are deformable to some degree), rigid body registration is performed in most of the articles [2, 12]. Application of nonlinear affine transformation on the whole floating images is a newer approach and having much more practical implementation [2,13, 14] in medical image

registration. The affine transformation preserves the parallelism of lines, but not their lengths or their angles.

Let  $\mathbf{T}$  denote the spatial transformation that maps features or coordinates from one image to another image. For 2-D affine registration the transformation matrix is:

$$\mathbf{x}' = \mathbf{a}^* \mathbf{x} + \mathbf{b}^* \mathbf{y} + \mathbf{c} \quad (1)$$

$$\mathbf{y}' = \mathbf{d}^* \mathbf{x} + \mathbf{e}^* \mathbf{y} + \mathbf{f} \quad (2)$$

or in matrix notation:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = T \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \quad (3)$$

where  $\mathbf{T}$  is a  $2 \times 3$  matrix of coefficients:

$$T = \begin{bmatrix} abc \\ def \end{bmatrix} \quad (4)$$

## 2.2. Computation of Similarity Metric

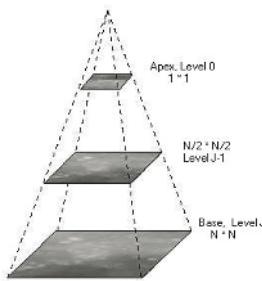
**Correlation Function:** We are matching multimodal images which are similar in some respects but dissimilar in other parts. But as long as there are sufficient similar structures in images, the matching algorithm performs well even with some dissimilarities present in either images. For affine transformation  $\mathbf{T}$  of floating image  $f$ , the correlation measure  $C_T$  of image  $f$  and reference image  $r$  is calculated using the formula,

$$C_T = \sum \sum r(\bar{x}) \times f(T(\bar{x})) \quad (5)$$

where  $r(x)$  denotes the intensity of the coordinates  $(x,y)$  in reference image  $r$ , the values of  $f(T(x))$  has been computed by affine transformation followed by bilinear interpolation of the gray values of the floating image  $f$ . If there is a match between  $r(x)$  and  $f(x)$ , the correlation of the two images will be maximum.

### 2.3. Multiresolution Approach

The fundamental theory of multiresolution imaging is to study the images at more than one resolution. A powerful but simple structure for representing the images at more than one resolution is the image pyramid. The base of the pyramid contains the highest resolution representation of the image; the apex contains a lowest resolution approximation. With moving up the pyramid, both size and resolution of the images decrease. In the present problem, we have used Haar wavelet transform to achieve multiresolution approach.



**Fig. 1.** A pyramidal image structure

Image pyramid: A powerful but simple structure for representing images at more than one resolution is the image pyramid. An image pyramid is a collection of decreasing resolution images arranged in the shape of a pyramid as shown in Fig. 1. The base of the pyramid contains the highest resolution representation of the image; the apex contains a low-resolution approximation. The level  $j-1$  approximation output is used to create approximation pyramid. The level  $j$  prediction residual output is used to build prediction residual pyramid. In present study we have applied two levels Haar wavelet transform. Initial information of affine transformation parameters are acquired from level  $j-1$  approximate image.

### 2.4. Optimization Technique

The application of a relatively new meta-heuristic strategy called *Particle Swarm Optimization (PSO)* has been presented in our study for medical image registration. The registration results show significant improvements compared to other optimization techniques like Genetic Algorithm as well as other techniques involving meta-heuristic optimization strategies, such as adaptive simulated annealing [18], tabu search method [19]. The original PSO algorithm introduced

by Kennedy and Eberhart [15] simulates the social behavior of a school of fish or a flock of birds, called the swarm. The individual swam members are called particles. Each particle remembers its own best position called the individual best. The best position found by the whole swarm is called the global best. Initially the algorithm is allowed to perform a very diverse search exploring a broad range of possible solutions. At each iteration a loop in the program is determined for each particle and the velocity of its neighboring particle is assigned to the velocity of the particle in focus. This simple rule creates a synchrony of movement. Unfortunately the population is quickly settled on a unanimous unchanging direction. Therefore a stochastic variable called craziness was introduced. At each iteration some change was added to randomly chosen velocities. This introduced enough variation into the system to give the simulation an interesting and lifelike appearance.

## 2.5 Proposed PSO Algorithm

In the registration problem let us consider the multimodality images to be registered are *Image A* and *Image B*. *Image A* is say reference image and *Image B* is the transformed image by affine transformation technique such that it will be correctly registered with *Image A*. Now for 2-D affine transformation, six parameters (equation 4) are required to transform an image. These six parameters are optimized by PSO, so that the image B transformed with the optimized parameters may be perfectly registered with the reference image (*Image A*). In every iteration, six transformation parameters to be optimized are updated by following two best values. The first one is ‘pbest’ the best value achieved so far by the particle. Another best value is ‘gbest’ obtained so far by any particle in the population.

1. In the reduced resolution, initialize the parameters randomly to obtain the rough idea of ‘pbest’ and ‘gbest’.
2. In actual resolution the parameters are again randomly distributed. Then update the parameter values with  $\text{present} = \text{present} + 2 \times \text{rand}() (\text{pbest}-\text{present}) + 2 \times \text{rand}() (\text{gbest}-\text{present})$   $\text{present} \rightarrow$  current value of the parameters in actual resolution
3. Calculate the fitness value (Objective function) using the updated values of the parameters.
4. If the fitness value is better than the best fitness value in the history, set current value of the parameters as the new ‘pbest’ value of the parameters.
5. Choose the objective function with the best value of all objective functions in the population and corresponding parameter values as ‘gbest’ value.
6. Stop when 30 iterations attained otherwise goto step 2.

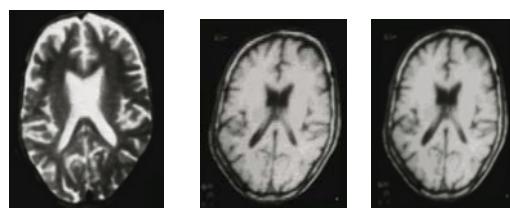
### **3 Experimental Results: Registration of Multimodality (CT & MR) Images of section of human brain having neurodegenerative diseases:**

In present work we have registered the ventricular region of section of human brain using CT and MR modalities. We have considered the ventricular region as region of interest ( ROI ) since the deformation of ventricular region in human brain indicates the stages of neuro degeneracy. Table 1 exhibits the improvement of correlation value ( $C_T$ ) before and after the registration process.

**Table. 1.** Illustration of the registration process for few sets of image data

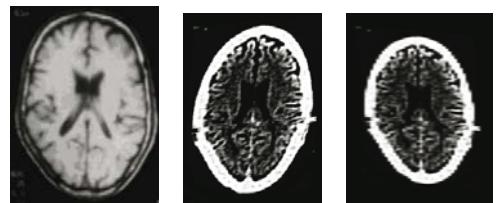
	$C_T$ before Registration	$C_T$ after Registration
SET-I	0.4811	0.5012
SET-II	0.5098	0.5715
SET-III	0.3773	0.5865
SET-IV	0.3185	0.3577
SET-V	0.7220	0.7308
SET-VI	0.4578	0.4631
SETVII	0.2520	0.3238
SETVIII	0.3961	0.4696

SET-I: MR  $T_2$  vs. MR  $T_1$  REGISTRATION



Reference Image      Floating Image      Registered Image

SET-III : MR  $T_1$  vs. CT Registration



Reference Image      Floating Image      Registered Image

SET-V: CT vs. MR T<sub>2</sub> Registration using PSO



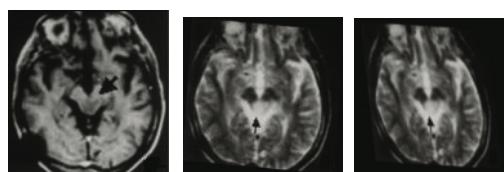
Reference Image      Floating Image      Registered Image

SET-VI: CT vs. MR T<sub>1</sub> Registration using PSO



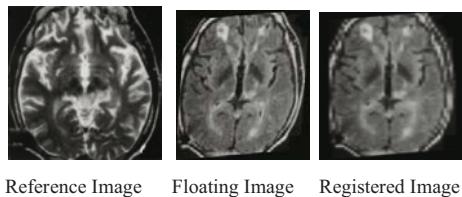
Reference Image      Floating Image      Registered Image

SET-VII: Preoperative MRI T<sub>1</sub> vs. Preoperative MRI T<sub>2</sub>



Reference Image      Floating Image      Registered Image

SET-VIII: Postoperative MRI T<sub>2</sub> vs. Postoperative MRI T<sub>1</sub>



## 4 Discussion

In biomedical image registration accuracy is the primary concern using appropriate transformation. Similarity metric functions used in registration problem are irregular, rough and often characterized by local optima. Most of the optimization techniques are accurate when initial orientation is much close to the transformation that yields the best registration. The approach to address this problem is to apply multiresolution techniques, whereby images are registered at increasing resolutions with initial orientations from preceding lower resolution. This method still frequently becomes trapped in local optima, as global optima may not present in lower resolution. Therefore a global optimization technique is necessary in medical image registration. Another important issue of optimization technique is the efficiency of the method. For this purpose the number of iterations required to achieve accurate registration must be as low as possible. However, an efficient global optimization technique is one, which provides a marked improvement in accuracy. It can be proved that PSO is noticeably more accurate over other evolutionary techniques.

## Acknowledgement

The authors would like to thank to Prof. Ravindranath, and Dr. P.K. Roy of National Brain Research Centre, Gurgaon, Govt, of India.

## References

1. Pluim J. P. W., Antoine Maintz J. B, Viergever M. A.: Mutual information based registration of medical images: a survey, *IEEE Trans. Medical Imaging* 22, 986–1004 (2003)
2. Bhattacharya M., Dutta Majumder D.: Registration of CT and MR images of Alzheimer's Patient: A Shape Theoretic Approach. *Pattern Recognition Letters* 21 (6 –7), 531–548 (2000)

3. Roche A., Pennec X., Malandain G., Ayache N.: Rigid Registration of 3-D Ultrasound With MR Images: A New Approach Combining Intensity and Gradient Information. *IEEE Trans. on Medical Imaging*, 20 (10), 1038–1049 (2001)
4. Comeau R. M., Sadikot A. F., Fenster A., Peters T. M.: Intraoperative ultrasound for guidance and tissue shift correction in image-guided surgery *Med. Phys.* 27 (4), 787–800 (2000)
5. van den Elsen P. A., Maintz J. B. A., Pol E.J.D., Viergever M. A.: Automatic registration of CT and MR brain images using correlation of geometrical features, *IEEE Transactions on medical images* 14(2):384–398 (1995)
6. Ritter N., Owens R., Cooper J., Eikelboom R. H., van Saarloos P. P.: Registration of stereo and temporal images of the retina. *IEEE Trans. on Medical Imaging*. 18(5), 404–418 (1999)
7. Wachowiak M. P., Smolíková R., Zheng Y., Zurada J.M., Elmaghhraby A.S.: An Approach to Multimodal Biomedical Image Registration Utilizing Particle Swarm Optimization. *IEEE Trans. On Evolutionary Computation* 8(3), 289–301 (2004)
8. Talbi H., Batouche M.: Hybrid Particle Swarm with Differential Evolution for Multimodal Image Registration *IEEE Int. Conference on Industrial Technology (ICIT)*, pp. 1567–1572 (2004)
9. Clerc M., Kennedy J.: The Particle Swarm—Explosion, Stability, and Convergence in a Multidimensional Complex Space *IEEE Trans. Evolutionary Computation* 6(1), 58–73 (2002)
10. McGuire M., Stone H. S.: Techniques for Multiresolution Image Registration in the Presence of Occlusions, *IEEE Trans Geoscience and Remote Sensing*. 38(3), 1476–1479 (2000)
11. Maes F., Vandermeulen D., Suetens P.: Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information *Medical Image Analysis*. 3(4), 373 –386 (1999)
12. Phan H. V., Lech M., Nguyen T. D.: Registration of 3D Range Images Using Particle Swarm Optimization *ASIAN 2004*, LNCS 3321, pp 223–235 (2004)
13. Bhattacharya M., Das A.: “Multi-Resolution Medical Image Registration Using Maximization of Mutual Information & Optimization by Genetic Algorithm”, Proc. of IEEE Nuclear Science Symposium/ Medical Imaging Conference (IEEE NSS/MIC-07), Hawaii, USA by IEEE Nuclear Science Society, pp. 2961–2964 (2007)
14. Bhattacharya M., Das A.: Affine Registration by Intensity and Fuzzy Gradient Based Correlation Maximization, accepted in proceedings of IEEE 7th International Symposium on Bioinformatics & Bioengineering (IEEE BIBE-07), 14–17 Oct. 2007, Boston, Massachusetts, USA (2007)
15. Kennedy J., Eberhart R.: Particle Swarm Optimization, *Proceedings of IEEE International Conference on Neural Networks*. 4, 1942–1948 (1995)
16. Kor S., Tiwary. U. S., Feature level Fusion of Multimodal Medical Images in lifting Wavelet Transform Domain, *Proc. of 26th Annual International conference of IEEE EMBS*, pp. 1479–1482
17. Bhattacharya M., Dutta Majumder D.: Knowledge Based Approach to Medical Image Processing. In *Pattern Directed Information Analysis (Algorithms, Architecture & Applications )*, publisher : New Age International Wiley, pp. 454–486 (2008)
18. Ritter N., Owens R., Cooper J., Eikelboom R. H., van Saarloos P. P.: Registration of stereo and temporal images of the retina, *IEEE Trans. on Medical Imaging*, 18(5), 404–418 (1999)
19. Chelouah R., Siarry P., Tabu Search Applied to Global Optimization, *European Journal of Operational Research* (123), 256–270

# **Relative Amplitude based Features of characteristic ECG-Peaks for Identification of Coronary Artery Disease**

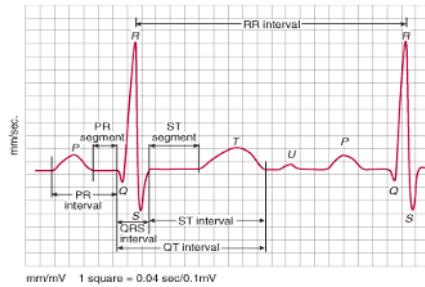
Bakul Gohel, U.S.Tiwary and T.Lahiri

Indian Institute of Information Technology, Allahabad. UP-India.  
{bcgohel, ust, tlahiri }@iiita.ac.in

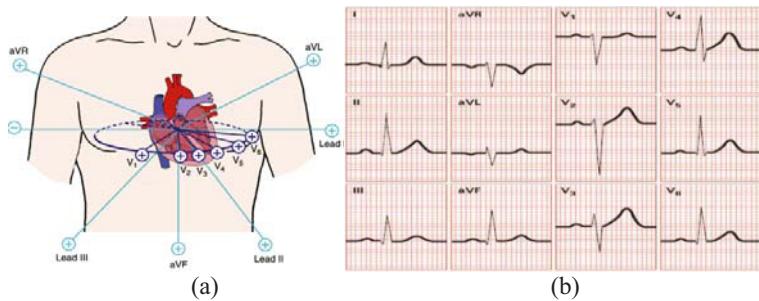
**Abstract:** Coronary artery disease or Myocardial Infarction is the leading cause of death and disability in the world. ECG is widely used as a cheap diagnostic tool for diagnosis of coronary artery disease but has low sensitivity with the present criteria based on ST-segment, T wave and Q wave changes. So to increase the sensitivity of the ECG we have introduced relative amplitude based new features of characteristic ‘R’ and ‘S’ ECG-peaks between two leads. Relative amplitude based features shows remarkable capability in discriminating Myocardial Infarction and Healthy pattern using backpropogation neural network classifier yield results with 81.82% sensitivity and 81.82 % specificity. Also relative amplitude might be an efficient method in minimizing the effect of body composition on ECG amplitude based features without use of any information from other than ECG

## **1 Introduction**

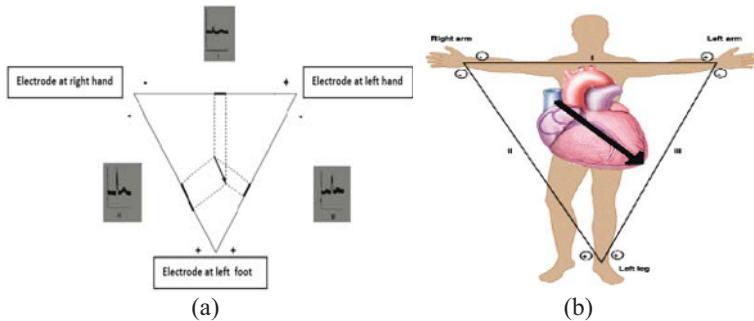
Coronary artery disease (Myocardial infarction) or Coronary heart disease is the leading cause of death and disability in the world [1, 8]. A successful development of diagnostic basis using ECG could be the cheapest diagnostic tool. Diagnosis of coronary artery disease from ECG mainly relies on ST segment elevation or depression, T wave inversion and prominent Q wave changes, but it has low Sensitivity in diagnosis of MI which range from 30% to 68% in various study [2, 3, 4, 8]. One more problem with ST segment and T wave based criteria is transient presence mean only present at time of acute heart attack. So there will be a need for finding the new features from ECG which has good sensitivity and specificity in recognizing coronary artery diseases and also has capability in estimating risk at early stage of disease.



**Fig. 1** ECG with characteristic waveform and time interval



**Fig. 2** Orientation of 12 leads (a) and normal waveform in each lead (b)



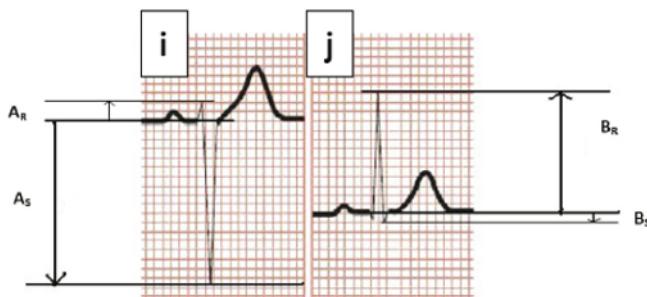
**Fig. 3** Orientation of cardiac vector and its projection on I, II, III limb leads (a) and electrode placement for I, II, III leads (b)

Recording fluctuation of electrical potentials of myocardial fibers (heart muscle) from body surface called the Electrocardiogram (ECG). Vector that

represents magnitude and direction of electric field generated by heart is called “net vector of polarization”. Amplitude recorded in any lead at given time is the projection of “net vector of polarization” on that lead at that time (Fig 2 and 3). The amplitude of the waveform recorded in any lead may be influenced by (1) Body impedance exert by intervening tissues (body composition) like fat, muscle, skin etc. [6]; and (2) the distance between the electrodes from heart [5]. Standard waveform of single cardiac cycle and 12 different leads of ECG are shown in Fig 1 and Fig 2b.

Our approach is based on the assumption that acute and chronic local ischemic effect or gross structural variation associated with CAD (like ventricular hypertrophy [8] ) affect the functionality of heart. These structural and functional changes cause deviation of “net vector of polarization or cardiac vector”, which in turn lead to change in amplitude of waveform in ECG. We try to identify this amplitude pattern of R and S waveform in different leads. But direct use of absolute amplitude is inappropriate due to effect of body composition on the ECG amplitude as mentioned above and it requires amplitude normalization. Methods for correction or normalization of amplitude of R and S wave use correlation analysis between ECG peak amplitude and one or more features of body composition like age, sex, body height, body weight, fat-free mass derived from bioelectric impedance method, skin fold thickness, cardiac mass in echocardiography, etc. [6, 9, 10]. In this paper, the concept of “Relative amplitude” has been introduced instead of using absolute amplitude. As it is a relative measure, it is assumed that it cancels out (or reduce) influence of body composition on ECG wave amplitude and get a normalized measure without requirement of information from other than ECG.

## 2 Relative Amplitude based features



**Fig. 4** Two ECG leads shows the amplitude of R and S waves

Relative Amplitude is the peak amplitude in one lead relative to the peak amplitude of another lead. It is a relative voltage difference of ‘R’ or ‘S’ wave, whichever maximum in one lead and ‘R’ or ‘S’ wave, whichever maximum in second lead. Example given below for calculating the “Relative amplitude” using lead i and j (Fig 4).  $A_R$ ,  $A_S$ ,  $B_R$ ,  $B_S$  is the average ‘R’ wave amplitude of first lead, average ‘S’ wave amplitude of first lead, average ‘R’ wave amplitude of second lead, and average ‘S’ wave amplitude of second lead respectively. Relative amplitude ( $L_{ij}$ ) between lead i and j can be found as

$$L_{ij} = (a_i - a_j) / (a_i + a_j), \quad i \neq j$$

$$a_i = \arg \max \{|A_R|, |A_S|\}$$

$$a_j = \arg \max \{|B_R|, |B_S|\}$$

Relative amplitude feature ( $\mathcal{V}_{ij}$ ) for one pair (i, j) is,  $\mathcal{V}_{ij} = [L_{ij} \ D_i \ D_j]$

$$D = \begin{cases} 1 & \text{if } R > S \\ -1 & \text{if } R < S \end{cases}$$

For the above example in Fig 4,  $\mathcal{V}_{ij} = [0.2 \ -1 \ 1]$

### 3 The Proposed Algorithm

1. Calculate average ‘R’ or ‘S’ wave amplitude whichever maximum from each lead of ECG.
2. Lead combinations and feature vector: here we used two lead combination from 6 limb leads (I, II, III, aVR, aVL, aVF) and chest leads (V<sub>1</sub>, V<sub>2</sub>, V<sub>3</sub>, V<sub>4</sub>, V<sub>5</sub>, V<sub>6</sub>) separately because distribution of fat on limb and on chest is not proportional, its vary with person to person. Relative amplitude features calculated for all combination mention below

15 combinations of Limb lead ( $\mathcal{V}_{ij}^L$ ): [ I II ], [ I III ], [ I aVR ], [ I aVL ], [ I aVF ], [ II III ], [ II aVR ], [ II aVL ], [ II aVF ], [ III aVR ], [ III aVL ], [ III aVF ], [ aVR aVL ], [ aVR aVF ], [ aVL aVF ]

15 combination of Chest lead ( $\mathcal{V}_{ij}^C$ ): [ V<sub>1</sub> V<sub>2</sub> ], [ V<sub>1</sub> V<sub>3</sub> ], [ V<sub>1</sub> V<sub>4</sub> ], [ V<sub>1</sub> V<sub>5</sub> ], [ V<sub>1</sub> V<sub>6</sub> ], [ V<sub>2</sub> V<sub>3</sub> ], [ V<sub>2</sub> V<sub>4</sub> ], [ V<sub>2</sub> V<sub>5</sub> ], [ V<sub>2</sub> V<sub>6</sub> ], [ V<sub>3</sub> V<sub>4</sub> ], [ V<sub>3</sub> V<sub>5</sub> ], [ V<sub>3</sub> V<sub>6</sub> ], [ V<sub>4</sub> V<sub>5</sub> ], [ V<sub>4</sub> V<sub>6</sub> ], [ V<sub>5</sub> V<sub>6</sub> ]

There are total 30 combinations and each combination has 3 Relative amplitude feature elements, so total length of feature vector ( $\mathbf{F}$ ) is  $30 \times 3 = 90$  for a one subject (12 leads ECG) (Table 1).

$$\mathbf{F} = [\dots \mathcal{V}_{ij}^L \dots \mathcal{V}_{ij}^C \dots]$$

**Table 1** Sample feature vector ( $\mathbf{F}$ ) for one subject

(i,j)	I,II	I, III	I, aVR	I, aVL	I, aVF	II, III	II, aVR	II, aVL	II, aVF	III, aVR
L <sub>ij</sub>	0.61	0.10	0.24	0.05	0.51	-0.54	-0.43	-0.58	-0.14	0.15
D <sub>i</sub>	1	1	1	1	1	1	1	1	1	-1
D <sub>j</sub>	1	-1	-1	1	-1	-1	-1	1	-1	-1

(i,j)	III,a VI	III,aVF	aVR,aVI	aVR,aVF	aVI,aVF	V <sub>1</sub> ,V <sub>2</sub>	V <sub>1</sub> ,V <sub>3</sub>	V <sub>1</sub> ,V <sub>4</sub>	V <sub>1</sub> ,V <sub>5</sub>	V <sub>1</sub> ,V <sub>6</sub>
L <sub>ij</sub>	-0.06	0.43	-0.20	0.13	0.47	0.01	0.05	-0.33	-0.23	0.00
D <sub>i</sub>	-1	-1	-1	-1	1	-1	-1	-1	-1	-1
D <sub>j</sub>	1	-1	1	-1	-1	-1	1	1	1	1

(i,j)	V <sub>2</sub> ,V <sub>3</sub>	V <sub>2</sub> ,V <sub>4</sub>	V <sub>2</sub> ,V <sub>5</sub>	V <sub>2</sub> ,V <sub>6</sub>	V <sub>3</sub> ,V <sub>4</sub>	V <sub>3</sub> ,V <sub>5</sub>	V <sub>3</sub> ,V <sub>6</sub>	V <sub>4</sub> ,V <sub>5</sub>	V <sub>4</sub> ,V <sub>6</sub>	V <sub>5</sub> ,V <sub>6</sub>
L <sub>ij</sub>	0.04	-0.33	-0.24	0.00	-0.37	-0.28	-0.04	0.10	0.33	0.23
D <sub>i</sub>	-1	-1	-1	-1	1	1	1	1	1	1
D <sub>j</sub>	1	1	1	1	1	1	1	1	1	1

3. Principle component analysis (*PCA*) has been used to reduce the redundancy so dimension of features. First 30 principle components with higher eigenvalue have been selected, so length of feature vector ( $\mathbf{F}$ ) reduced to 30 from 90.
4. Backpropogation neural network has been used for classification of MI and healthy control (HC) class because it has high tolerance to noisy data, ability to deal non-linear classifiable problem. Architecture of backpropogation neural network used as follow.

Input layer: 30 tansig neuron

Hidden layer: 6 tansig neuron

Output layer: 1 purelin neuron

Training Function: Gradient descent

Performance Function: Minimum Square Error (MSE)

#### 4 Results and Discussion

We used digitized 12 leads ECG signals of myocardial infarction and healthy control with mean age 57.2 for man and 55.5 for female from “PTB diagnostic ECG database”. ECG signal was digitized with 1000 Hz sampling rate [7]. The backpropagation neural network has been used for classification of MI and HC ECG data using relative amplitude based features from ECG as mentioned above. We have chosen the training dataset (60 MI and 30 HC) and testing datasets (44 MI and 22 HC) randomly from database, with best result. For computation we used MATLAB. We got the following results

$$\text{Sensitivity} = (\text{TP}) / (\text{TP} + \text{FP})$$

$$\text{Specificity} = (\text{TN}) / (\text{TN} + \text{FN})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

Where TP, FP, TN and FN are true positive, false positive, true negative and false negative respectively.

**Table 2** Training and testing ECG data set (From PTB diagnostic ECG database)

Training data set			Testing data set		
Total	MI	HC	Total	MI	HC
90	60	30	66	44	22

**Table 3** Result of backpropagation neural network classification using relative amplitude based features

TP	FN	TN	FP	sensitivity	specificity	accuracy
36	8	18	4	81.82 %	81.82 %	81.82%

Relative amplitude based features of 12 lead represent structural and functional changes of heart shows capability in differentiating MI and HC class with sensitivity of 81.82 % and specificity of 81.82%. True sensitivity and specificity of these features only can be estimate by applying on dataset labeled with some confirmatory diagnosis apart from ST, T and Q wave criteria.

## 5 Conclusion and Future Work

It has been shown that relative amplitude based features show remarkable capability in discriminating Myocardial Infarction and Healthy pattern using backpropagation neural network. In this paper, relative amplitude based features are used as efficient tool for screening out coronary artery disease. However, Relative amplitude might be an efficient method in minimizing the effect of body composition on ECG amplitude based features without use of any information from other than ECG and the resultant features can be used for other applications. Combining relative amplitude features with other parameters like ST segment and T wave changes might be further increase the sensitivity and specificity of ECG in diagnosis of myocardial infarction.

## References

1. Latheef S.A.A., Subramanyam G.: Prevalence of Coronary Artery Disease and Coronary Risk Factors in an Urban Population of Tirupati. Indian Heart J 59, 157–164 (2006)
2. Ginn P. H., Barbara J.: How accurate is the use of ECGs in the diagnosis of myocardial infarct?. Journal of Family Practice 55(6), 539–540 (2006)
3. Asch F.M., Shah S., Rattin C., Swaminathan S., Fuisz A., Lindsay J.: Lack of sensitivity of the electrocardiogram for detection of old myocardial infarction: a cardiac magnetic resonance imaging study. Am Heart J. 152(4), 611–612 (2006)
4. Zimmerman J., Fromm R., Meyer D., Boudreaux A., Chuan-Chuan C., Smalling R., Davis B., Habib G., Roberts R.: Diagnostic Marker Cooperative Study for the Diagnosis of Myocardial Infarction. Circulation 99, 1671–1677 (1999)
5. Edhouse J., Brady W.J., Morris F.: ABC of clinical ectrocardiography. BMJ 324, 963–966 (2002)
6. Tochikubo O., Miyajima E., Shigemasa T., Ishii M.: Relation Between Body Fat-Corrected ECG Voltage and Ambulatory Blood Pressure in Patients With Essential Hypertension. Hypertension 33, 1159–1163 (1999)
7. PTB diagnostic ECG database , <http://physionet.org/physiobank/database/ptbdb>
8. Current Medical Diagnosis & Treatment 2007(CMDT), chapter 05 cardiology, McGraw Hill's AccessMedicine
9. Okin P. M., Jern S., Devereux R. B., Kjeldsen S. E., Dahlöf B.: Effect of Obesity on Electrocardiographic Left Ventricular Hypertrophy in Hypertensive Patients. Hypertension 35, 13–18 (2000)
10. Lukaski H. C., Bolonchuk W. W., Hall C. B., Siders W. A.: Validation of tetrapolar bioelectrical impedance method to assess human body composition. J Appl Physiol 60, 1327–1332 (1986)

# ProVis: An Anaglyph based Visualization Tool for Protein Molecules<sup>1</sup>

Rajesh Bhasin and Abhishek Kumar

Computer Science and Information Systems Department,  
Birla Institute of Technology and Science – Pilani, Goa Campus, Goa, India  
{rajesh045, abhishek.kumar.ak}@gmail.com

**Abstract.** Proteins are highly complex and flexible structures. Rendering tools are often used to study their dynamics, functions and in some cases even deformations that arise due to slow dynamics. In this work, we present ProVis, a visualization tool for 3-dimensional rendering of protein molecules with the ability to create and display anaglyph images. The tool allows for viewing the protein molecules, locating atoms, viewing bonds, viewing the protein backbone, searching for specific bonds, visualizing them and also analyzing the various scalar properties of the protein. The use of display lists allows for very fast rendering and real time animation of the complex molecules with negligible latency time between scenes. The novelty of the tool is the interface which combines the power of 3D visualization using anaglyphs with the geometry and graphics of the protein to provide a real-time interaction environment for large amounts of abstract data.

## 1 Introduction

**Motivation.** Biologists have long felt a need for an interface that not only lets them analyze protein molecules, but is also comfortable to use and provides information about the various structures present in the protein. Analysis of proteins involves efficient visualization of these structures and its residual constituents along with a mapping of the different scalar properties along its surface. The motivation behind our work was primarily to assist biologists with a tool that could be used for selecting, transforming and representing abstract data in a form that facilitates human interaction for exploration and understanding. ProVis is also aimed at assisting biologists in analyzing protein-protein interactions.

**Related Work and Our Approach.** While rudimentary molecule structures have been previously explored, a large number of data sets of the 3D structure of macromolecules have only become available in recent years. They are archived in the Protein Data Bank [1]. VMD [2], PyMOL [3], and MolScript [4] are some popular

---

<sup>1</sup> This work was done as part of summer internship at the Department of Computer Science and Automation, Indian Institute of Sciences, Bangalore

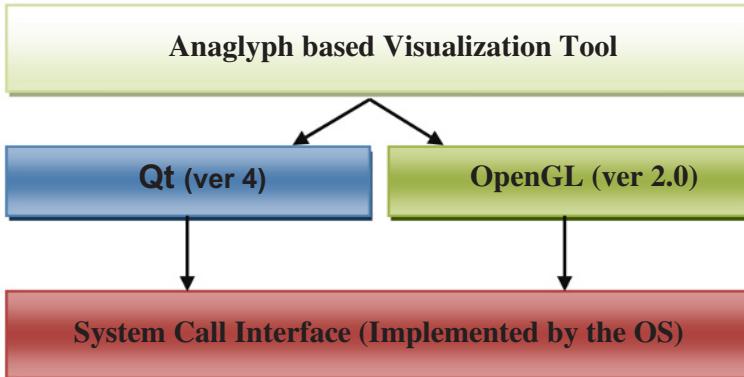
software packages for molecule visualization. These, along with OpenBabel[5], BALLView[6] and Python Molecular Viewer (PMV)[7] provide elementary visualization capabilities. Avogadro[8] and Kalzium(as a part of the KDE education project)[9] extend OpenBabel to incorporate various other features such as distance measurement, angle measurement and geometry optimization. RasMol[10] provides powerful display features, but the user interface is limited to the command line. Although RasTop[11] provides a graphical-user-interface wrapper to RasMol, both these tools lack the ability to augment the display with additional information, such as change in surface texture, visualization of various parameters like temperature, hydrophobicity etc. Protein Explorer[12] provides almost all the features of RasTop in addition to being more user-friendly. AISMIG[13] is a server side Molecule Image Generator which displays images of protein files uploaded by clients. Not only does this tool suffer from lack of interactivity, but the only form of visualization is through rendered images. Molecule Spatioplotter[14] provides an elementary display tool with support for user-defined file formats. Santorini[15] allows for animation, surface calculation and computation of specific distribution functions along with simple visualization.

In addition to the basic visualization capabilities provided by these tools, ProVis allows for the use of anaglyph images for perception of depth. The user may also augment the displayed structure with a 3D scalar field according to a distribution (such as electron complementarity or hydrophobicity), for better visualization of the properties of the molecular structure. Further, we also provide support for speech recognition, implemented by using a speech recognizer API.

This paper has been divided into three parts viz. Architecture, Application description and Interface design. The Architecture part deals with the system description; Application description discusses the methods we have used for the perception of the structures and shapes present in the molecule, and an overview of the visualization process. The Interface design section talks about the design and ergonomics. Towards the end we discuss the various enhancements that may be added to the software tool to increase usability and interactivity. Here, we have used protein 1DKF as a test input which contains approximately 4500 atoms.

## 2 Architecture

We present a tool which loads different files in the standard “.obj” geometry definition file format[16] containing data about the protein regarding structures, scalars and other attributes, and generates a 3-D visualization of this data as the output with real time interaction capabilities. One of the most important parts of a visualization tool is the architecture that supports it. The underlying architecture developed here is based upon the principle of graphics pipeline[17].

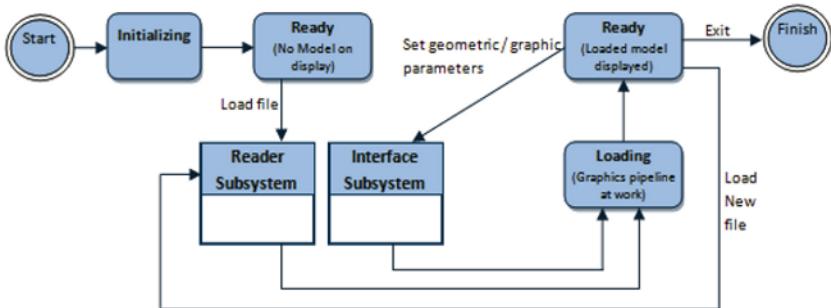


**Fig. 1(a).** Architecture Overview

The abstract data containing information about the model is passed through the pipeline before it is rendered onto the screen. Abstract data as such is of little use. However, if the abstract data is visualized and shown, its understandability increases manifold and it can help scientists in comprehending complex processes that were not clearly visible earlier. The pipeline essentially involves constructing objects from the abstract data in their own coordinate system using OpenGL[18] primitives and then positioning them within a global coordinate system.

Figure 1(a) shows the various components of the visualization tool, while 1(b) provides an overview of the functioning of ProVis through a state transition diagram. The arrows indicate dependencies among the various components. The OpenGL[18] library was used to implement the graphics and rendering, while Qt[19] was used as the front-end application development framework since it runs on all major platforms and has extensive internationalization support.

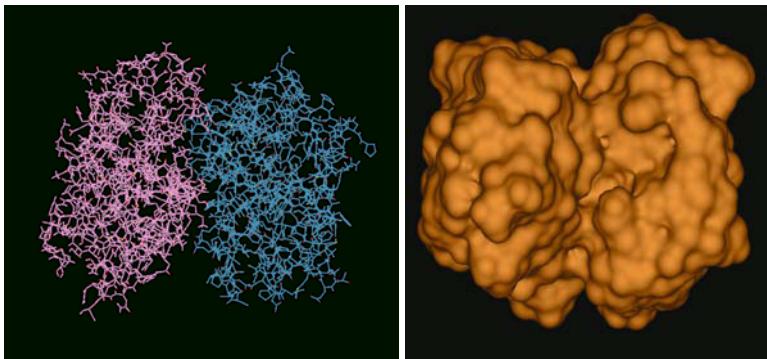
The application comprises three modules – the interface, framework and reader modules. The interface module is responsible for loading and displaying the graphical user interface, the framework module stores information about the various structures and representations for storing the protein molecule data. The reader is responsible for reading the data files and interpreting them appropriately.



**Fig. 1(b).** State Transition Diagram (overview) of the system.

### 3 Application Description

In order to improve the dynamics of the visual representation we have used Display Lists which compile and store the abstract data as primitives (opengl commands) in the memory space of the graphics card(s). As a result a significant speedup is observed in the rendering process. Once built, only few additional operations need to be applied to the display lists before it is displayed. The test input used (protein 1DKF) has 66402 triangles and with the use of display lists we were able to achieve real time interaction and animation.



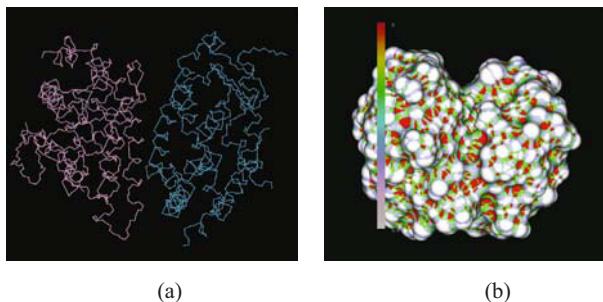
**Fig. 2.** Protein 1DKF displayed (a) without (b) with an enclosing surface

**Modes of display.** The application supports the visualization of protein in different modes viz. atoms, bonds, alpha shapes and surfaces. Each of these is also shown in 3D using anaglyphs. There are two modes of representation: the atoms and backbone. In both these modes, segregation is done based on the chains present in the structure. Atoms and bonds which do not belong to the main chain of bonds (also called the backbone) and are part of residual groups are also shown separately. The algorithm

used also visualizes different alpha shapes like tetrahedron, triangles and edges by drawing the surfaces appropriately. We use the software implementation of a trackball to provide a natural interface to the user, where virtual objects are handled like real life objects. In order to improve the sense of depth, lighting is added to the scene using the APIs provided by OpenGL. This improves the visibility and the user can easily analyze the finer details on the surface of the protein molecules.

**Structure and display features.** Multiple renderings of an object from various angles and viewpoints may lead to new information. It also provides a better understanding of how the shape determines its function, reactivity and other such factors. Such relations can best be explored by generating high quality 3D visualizations and providing an intuitive navigation interface which allows for an in-depth inspection of specific details. We incorporate this feature by allowing the user to see all the structures present in the protein simultaneously and still have access to relevant details of one particular form or property (such as the locations of C-C bonds). Hence, one of the unique advantages of using our interactive and feature-rich application is that user can quickly try different approaches to visualizing data.

Detailed wireframe-like representations show the atomic structure, including information on size, shape, and surface. These models, combined with different material colors, provide a better understanding of the underlying structure. Since the different models emphasize diverse aspects of the structure, a visualization of different representations is critical for the comprehensive analysis of the various features of the shape, structure and the surface of the protein molecule.



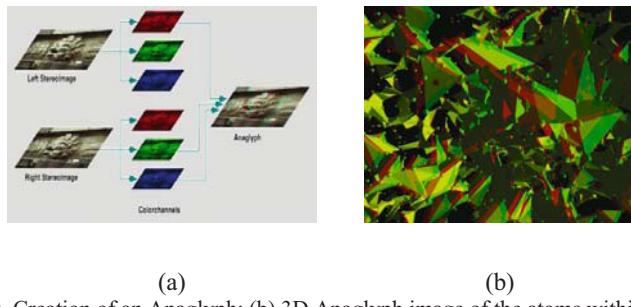
**Fig. 3.** (a) Protein 1DKF's backbone view; (b) Visualization of a field using scalar mapping on the protein surface

**Zoom.** For an insight into the structure of the protein, it is essential to be able to rotate freely and ‘zoom’ the structure from different angles.

**Scalars.** The viewer also visualizes various scalars associated with the protein surface (Fig 3b). The scalar values are linearly interpolated and then mapped onto the surface. These scalar values are calculated based on mathematical functions and provide a means to visualize various fields or temperature, pressure gradients etc on the surface of the molecule. The colour legend is also shown along with the protein surface and scalar values associated with different colours are shown alongside. This

helps the user to correlate the values along the surface with the colours used while he interacts with the protein simultaneously. Both, the colour and the mapping mechanism are entirely user defined and may be used to map any feature that needs to be studied.

**Anaglyphs.** The strongest feature incorporated in the viewer is the use of anaglyphs to give a better perception of depth. We draw two images of the same protein by changing the position of the camera and then superimpose them using the accumulation buffer. Two different colors are used for drawing the two objects. These images can then be viewed using 3D stereographic glasses, which have different colors for the left and the right eye. As a result they filter out one image out of the two and each eye sees only one image (which is of different colour). When images fall on the back of the retina of a human eye, they are superimposed by the brain to give a perception of depth in the structural representation of the protein. This is particularly useful while studying the different cavities present in a protein. These cavities are active sites for protein interactions and hence form an important part in the study of proteins. In Figure 4b, we see an anaglyph image of the internal structure of a protein molecule.



**Fig. 4(a).** Creation of an Anaglyph; (b) 3D Anaglyph image of the atoms within a 1DKF protein molecule along with alpha shapes.

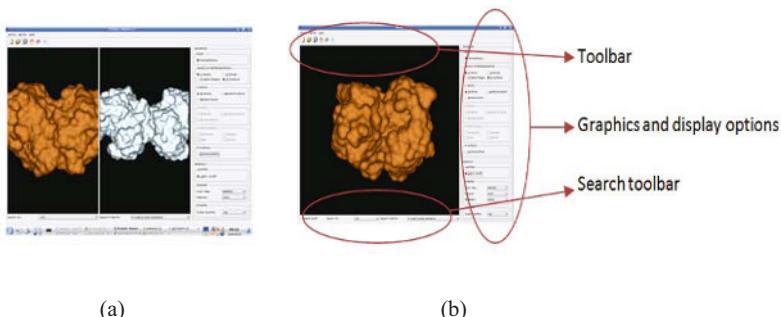
**Animation.** The visualization tool provides for animation of the protein where its position is rotated by a few degrees over small intervals of time. The user can change the speed and the direction of the rotation by using the trackball and can see the updated position in real time. It is also noteworthy that while the animation is in progress, the user may switch between the modes, change colors and set or tune the various graphic parameters including light and material colors.

## 4 Interface Design

One of the most important issues in the design of any interface is its simplicity and usability. Shneiderman's "Eight Golden Rules of Interface Design" [20] proposes a collection of principles, which is derived heuristically from experience and applicable in most interactive systems. The eighth principle stresses the importance of reducing the short term memory load on the user. In this respect, our application presents a simple user interface with minimal and requisite buttons. The other necessary options appear only when the requirement arises.

All the utilities are kept together in one column on the right-most part of the GUI for easier access. The GUI is provided with a toolbar from where the user can select the mode of representation. A status bar displays a description whenever the mouse is pointed over a toolbar item. The design of the interface is hierarchically divided into (1)Geometry, and (2)Graphics. Since the interface is hierarchically designed, further additions to the system are easy. The various viewing options are displayed in the geometry section of the interface. The graphics section handles various graphics aspects including lighting and material colours. These attributes, when suitably selected, provide an enhanced view of the protein. The toolbar below the viewer allows the user to search for specific bonds within the structure, which then get highlighted for better view and visualization.

The region where the molecule is rendered or displayed, also called the viewer, has two modes of operation: (1) normal (the default), and (2) stereo (available by toggling the appropriate option from the right column). Since we use anaglyphs to create stereoscopic vision, one can use a 2-colour glass (each lens a chromatically opposite color, usually red and cyan) for viewing them.



**Fig. 5.** (a) Two-pane view of the interface;  
 (b) A description of the different parts of the graphical user interface.

Two other unique features of the interface include two pane viewing mode, which allows for viewing two different protein molecules at the same time, and speech recognition, which allows the user to issue commands using speech, thereby providing an interface accessible to the physically-disabled or handicapped.

## 5 Conclusion

We present a visualization tool which is capable of loading and displaying the various protein molecules available from the protein data bank. The tool allows for viewing the protein molecules, locating atoms, viewing bonds, viewing the protein backbone, searching for specific bonds, visualizing them and also analyzing the various scalar properties of the protein.

In addition to providing a simple, intuitive and powerful interface for viewing the complex structure of protein molecules, the tool also allows for very fast rendering and animation along with specialized graphics features such as lighting which provides a real-time interaction environment for the user to visualize the structure. The ability to create an anaglyph image allows the molecule to be perceived in 3D depth by the use of 2-colour glasses which are an inexpensive replacement for other costly software that rely on virtual reality equipment for similar visualization requirements.

## References

1. Berman, H., Westbrook J., Feng Z., Bhat G., Weissig H., Shindyalov N., Bourne P. E.: The Protein Data Bank. *Nucleic Acids Research* 28, 1, 235–242, (2000)
2. Humphrey W., Dalke A., Schulter K.: VMD - Visual Molecular Dynamics. In *Journal of Molecular Graphics* 14, pp. 33–38, (1996)
3. Delano L.: PyMOL. In DeLano Scientific LCC, (1998–2004)
4. Kraulis J.: MOLSCRIPT - A Program to Produce Both Detailed and Schematic Plots of Protein Structures. In *Journal of Applied Crystallography* 24, pp. 946–950, (1991)
5. OpenBabel: The Open Source Chemistry Toolbox, <http://openbabel.org/>
6. BALLView: Molecular Visualization Application, <http://www.ballview.org/>
7. PMV: Python Molecular Viewer, <http://mgltools.scripps.edu/>
8. Avogadro: Advanced Molecular Editor, <http://avogadro.openmolecules.net/>
9. Kalzium: Visualization Software, <http://edu.kde.org/kalzium/>
10. Sayle A., Milner-White J.: RASMOL: Biomolecular Graphics for All. In: *Biochemistry* 20, 374–376, 1995
11. RasTop: Molecular Visualization Software, <http://www.geneinfinity.org/rastop/>
12. Protein Explorer: Protein Visualization Software, <http://www.proteinexplorer.org/>
13. Bohne-Lang A., Groch W., Ranzinger R.: AISMG-An Interactive Server-side Molecule Image Generator. In *Journal of Nucleic Acid Research*, Vol 33 (Web Server issue) (2005)
14. Molecule Spatioplotter-3D Molecule Viewer, <http://www.iksoft.com/proj/chemistry/molecule.php>
15. Santorini: Molecular Viewer, <http://development.oeffner.net/>

16. Wavefront .obj file format specification for the Advanced Visualizer software, <http://local.wasp.uwa.edu.au/~pbourke/dataformats/obj/>
17. Lampe O., Viola I., Reuter N., Hauser H. : Two -level Approach to efficient Visualization Of Protein Dynamics. In IEEE transactions on Visualization and Computer Graphics, 13 (6), 1616–1623 (2007)
18. The OpenGL API for graphics, <http://www.opengl.org/>
19. The Qt application development framework, <http://trolltech.com/products/qt/>
20. Schneiderman B., Plaisant C.: Designing the User Interface: Strategies for
21. Effective Human-Computer Interaction. Addison-Wesley Publishing Company (1987)

# Applying Cognitive Psychology to User Interfaces

Sabeen Durrani and Qaiser S. Durrani

FAST – NU, Computer Science Department Lahore, Pakistan  
sabeen.durrani@lhr.nu.edu.pk, qaiser.durrani@nu.edu.pk

**Abstract.** This paper explores some key aspects of cognitive psychology that may be mapped onto user interfaces. Major focus in existing user interface guidelines is on consistency, simplicity, feedback, system messages, display issues, navigation, colors, graphics, visibility and error prevention [8–10]. These guidelines are effective in designing user interfaces. However, these guidelines do not handle the issues that may arise due to the innate structure of human brain and human limitations. For example, where to place graphics on the screen so that user can easily process them and what kind of background should be given on the screen according to the limitation of human motor system. In this paper we have collected some available guidelines from the area of cognitive psychology [1, 5, 7]. In addition, we have extracted few guidelines from theories and studies of cognitive psychology [3, 11] which may be mapped to user interfaces.

## 1 Introduction

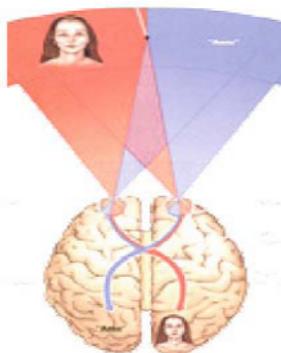
While designing an interface one of the major goals of a designer is to design a system which can be used by the users effectively, even novices, without spending much time in understanding how to use the system. To accomplish this goal user interfaces should be designed according to human capabilities and limitations. Extensive research material is available regarding guidelines for user interface design. But limited number of these guidelines is based on cognitive aspects of humans. These cognitive psychology based guidelines are generally about human memory limitations, attention, learning, decision making and perception [7–9]. However, there are many other theories and studies of cognitive psychology that provides us with guidelines to design better user interfaces. Such theories and studies are organization of brain [3, 11], Simon effect [6], transference [5], mental imagery [3, 11], limitations of motor system [3], interference [3], visual search [3] and cognitive abilities [3]. Limited research material is available that map these studies of cognitive psychology to user interface design [1, 5]. In this paper a set of guidelines is presented based on the theories of cognitive psychology. These guidelines will help an interface designer in designing interfaces that may reduce cognitive load on the user by presenting information in a more understandable

manner. There are many issues that need addressing in interface design including the following: text and image location on the screen, background and text colors, information presentation, placement of the most important information, handling cognitive abilities of users and keeping user involved in the system.

## 2 Guidelines

### 2.1 Placement of Text and Images

Here we discuss the issue of placing text and images on the screen so that the user pays attention to them. To resolve this, we have to understand the structure of human brain to see how it pays attention to text and images and which portions of the brain are allocated for their processing. The brain is divided into two hemispheres; each of them is specialized for different types of processing. The *left hemisphere* is associated with linguistic and analytic processing, whereas the *right hemisphere* is associated with perceptual and spatial processing. Brain processes information contralaterally; that is left hemisphere controls right side of the body while right hemisphere controls the left side of the body. Hence, left eye is controlled by the right hemisphere and right eye is controlled by the left hemisphere [3]. Hence, user can identify *text* quickly when they are displayed in the user's *right visual field* and *images* are identified quickly when presented in the *left visual field*, as shown in figure 1 [1].



**Fig. 1.** Visual Fields [1]

Understanding the brain structure helps us in positing text and images on screen. Text should be placed so that it falls in the user's right visual field and images should be placed where it will fall in the user's left visual field. This will facilitate the processing of text and images in the corresponding specialized hemispheres and will help reducing the cognitive load on the user.

## 2.2 Simon Effect

According to Simon effect [6] *people respond faster and more accurate to stimuli that occur in the same relative location as the response*, even if the location of information is irrelevant to the actual task. This is because there is an innate tendency in humans to respond towards the source of stimulus. In this context when location of stimuli and response is same, this is called **congruent condition** and opposite to this condition is called **incongruent**. Simon effect deals with congruent condition. For example, if a pilot is flying a plane and the left engine has a problem, the indicator for the left engine should be positioned to the left of the indicator, for the right engine. If it is the other way around, the pilot may not respond correctly to the indicator and adjust the wrong engine. That could be problematic.

To understand how this finding can help in computer interface design, consider the following scenario: The task is to press button1 if an image appears on right half of screen and press button2 if image appears on left half of the screen. Here image is stimulus and button is its response. Buttons should be placed at the same location where the corresponding image will be displayed. That is button1 on right side of screen and button2 on left side of screen. This will help user in quickly pressing the button based on the appearance of corresponding image. Hence *interface should be designed to provide congruent condition in the context of Simon effect.*

## 2.3 Role of past knowledge of user

Past knowledge plays an important role in understanding new concepts and increases the perceiving ability. Transference and mental imagery are important concepts in *role of past knowledge of user* in designing interfaces.

### 2.3.1 Transference

“Transference refers to the expectations of a user about an interface’s behavior based on his/her previous experiences with other interfaces [5]”. This may include layout design, placement of buttons, shape and functionality of buttons, navigation bar designs, functionality of links, meaning of labels and icons and working of scroll bar. When an interface is according to the expectations of a user, the effect is *positive transference*. On contrary, when an interface does not map on to the expectations of user, the effect is *negative transference* and as a result user may commit mistakes during their interaction with the interface. *Designers should make interfaces that result in positive transference in order to make the system easy to use for a user.*

### 2.3.2 Mental Imagery

The mental representation of how things look is referred to as *mental images* [3]. While making interfaces designers should keep in mind that every thing must be in accordance with mental images of the user to facilitate his/her perception and hence processing. S/he should not spend time in understanding the things that s/he did not encounter in their lives. For example, users of Microsoft office have the mental image about the save button that looks like this:  They may have difficulty in finding save button when they work on open office where save button looks like this:  Although the basic icon is a floppy but the shape is different and this may create problems for a novice user. Similarly, users of internet explorer face difficulty when they start using Mozilla Firefox. One example is the two different terms used in these two browsers for the same task. Internet explorer uses term “refresh page” with this icon  while Mozilla Firefox uses term “reload this page” with this icon  . If we compare the layout of the two web browsers shown in figure 2, we will find that most of the icons used in these two are based on the same idea or sketch but the appearance is a little different that puts cognitive load on the user.



(a)



(b)

**Fig. 2.** Interfaces of two well known browsers. (a): Mozilla Firefox; (b): Internet Explorer.

We would suggest that *if the purpose of a set of interfaces is same then the layout should not be changed*. For the case discussed above, the basic purpose of the two applications is to help the user in web browsing, hence the interface should be the same so that user does not get confused.

Concept of mental imagery is very important in the case of educational systems. *Examples, images and information used in an interface of an educational system should be based on previous knowledge of the user*, rather than introducing the concepts that the user is unaware of. This would facilitate the learning process of a user. For example, if an interface is to be designed to teach English alphabets to the children then those images which are familiar to them should be chosen. The previous knowledge is based on the culture and society of the child. For example, a child in Africa is familiar with different things as compared to a child in Asia (the things that are specific to a culture).

Previous knowledge depends upon many factors including culture, traditions, family background, educational background, etc. For example if a user wants to have information about working of human heart and he has no background in medical science, the information presentation and organization for this user should be different from the user having medical background.

#### **2.4 Limitation of Motor System**

*Humans feel difficulty in getting one motor system to do two things at once.* There are bottlenecks in the auditory and visual systems: the points at which one can attend to only one spoken message or one visual image at a time [3]. While designing an interface designers should take care of this limitation of human motor system. For example designer should not give background music if some audio information is already being given to the user. Some designers think that while giving background music they will prevent their users from boredom, but they don't realize that by doing so they are putting extra cognitive load on their users. In case of visual information the same guideline applies but replaces background music with background image. When visual system is involved in reading and understanding the text then it should not be distracted by the background image. Hence, background images should be avoided, especially those containing text. Following image further clarifies this point where background text is interfering with main text written in foreground:



**Fig. 3.** Interface where background text is interfering with foreground text [12]

This is also a common experience that one cannot pay equal attention to both, the text s/he is reading and the music s/he is listening to, because attention is divided into two tasks, hence decreasing the performance. This fact should be taken into account while designing interfaces and one should avoid background music if the sole purpose is to provide information. This is especially true in case of learning systems. However this also depends upon the user group of the system because there are people in the world who grab information more quickly with music, so user preferences should be identified before applying this guideline.

## 2.5 Memory Limitation

Short term memory has a limited capacity to hold and process information [3]. Based on different experiments this capacity is generally found to be  $7\pm2$  items or chunks of information. *While displaying information the memory capacity of  $7\pm2$  should be kept in mind.* For example, number of concepts introduced to the readers at a time should not exceed  $7\pm2$ . In order to increase retention of user, information displayed may be divided into  $7\pm2$  chunks on some meaningful basis [2]. *Interface designers should keep number of steps small to accomplish a certain task.* Interface should be designed in a way to reduce memory load on users by minimizing the need of memorizing things [7].

## 2.6 Human perception

Perception is the ability to recognize and identify information from the environment. There is a need to facilitate perception of users, so that they perceive the required information timely and correctly. In order to do so, *icons on graphical user interface should convey their meaning to the users. Sounds and audio should be audible and convey the message clearly. Text should be visible and its color should not conflict with the background*, i.e., text and background color should be chosen to assist user in reading [7].

Readers make mental associations among different objects (text or images) depending upon the visual cues present on the screen. *Proximity* and *Similarity* are the two rules suggested by Gestalt psychologist. These concepts are briefly explained below [5]:

### 2.6.1 Proximity

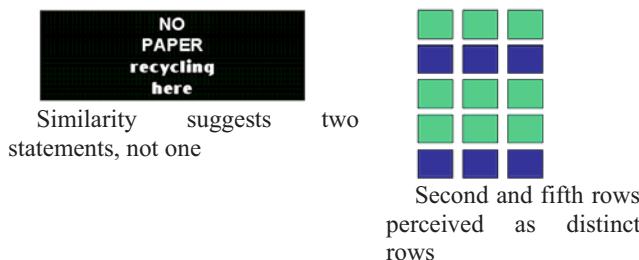
Rule of proximity indicates that items close together are perceived as being related or associated.



**Fig. 4** Rule of Proximity [5]

### 2.6.2 Similarity

Rule of similarity indicates that items with a similar appearance are perceived as being related or associated:



**Fig. 5** Rule of Similarity [5]

Hence *rules of proximity and similarity should be kept in mind while designing interface, to prevent readers from making wrong relationships between objects*. For example, in **navigation bar design**, its items should be proximal and similar so that users perceive them as objects being together.

Search is an activity that humans need to perform whenever they look for some relevant information. Visual search is an important concept that needs to be catered properly in user interface design.

### 2.6.3 Visual Search

User has to spend a lot of time to search some specific information from a bulk. Sometimes, user may not be able to find the required information although that information is present. This is because the information is not categorized into appropriate categories. According to cognitive psychology, humans can work easily with information presented in categories [3]. *Information should be categorized in a way to facilitate visual search of the users [3]*. Categories should be made such that the relevant information is in one category with the most appropriate label of the category. The label should not be in conflict or ambiguity with any other category label. A category label should be chosen so that it fully depicts the main theme of information and multiple users should not perceive different meanings from it.

Search can be facilitated through hints, for example, if we have to find a house of a friend in a street and if we know that the house is built of red bricks then we can easily find the house [3]. This finding of cognitive psychology should be applied to information categorization as well. Again *category labels should be chosen such that they can be served as hints for the users to find relevant pieces of information from bulk of information*.

## 2.7 Learning

*Interface should be designed such that users can learn quickly how to use it.* Interface should adapt according to each user by providing only basic functionalities to novices and advanced features only to the experienced users [8]. Many applications have been developed to teach users different topics and concepts. Interfaces of such applications should facilitate learning of users by presenting interface according to a user's limitations, strengths and requirements.

## 2.8 Role of Unusual Things – Minimize Interference

According to cognitive psychology, humans pay more attention to some things that are unusual [3]. *Designers should take care of not displaying unimportant information (textual or pictorial) in attractive (funky, glittery, animated etc.) form, because this may distract the user from paying attention to the important information.* For example, if the purpose of the system is to give information to children regarding brain working then there is no need to display animated cartoon images just to make the system attractive to the child. Rather, the images and information should be relevant to the brain's working mechanisms. However, those images should be designed in a manner that would help the child in getting and understanding the information in an enjoyable way.

Interference can be in the form of unusual and unimportant things presented in a way that may distract the user from the original task that s/he is performing. This interference should be avoided especially in case of educational or information systems.

## 2.9 Which Colours to Choose?

According to Shneiderman colors soothe or strike to the eye and play an important role in identification and recognition of objects [8]. They increase retention of objects by 50% [4]. This emphasizes the importance of choosing appropriate colors for background, foreground, images and text. Number of colors used in an interface should be decided according to the human capabilities of processing them. Colors chosen should enhance attention, recognition, interest and understanding of a concept. This all seems like a challenge but cognitive psychology helps us in addressing this challenge.

### 2.9.1 Choose minimum number of Colours

Humans can store five to nine elements in their short term memory and this fact can also be applied on the number of colors used in an interface. If we want to present many related pieces of information, we can use the same color for similar information, so that user can easily identify the relationship between alike information, i.e., color coding [9]. It is important to *limit the number of colors to 7±2 in interfaces* used as in the above kind of scenarios [8] [10]. This helps in reducing the cognitive load on the user.

### 2.9.2 Avoid using too many Bright Colours

Usually designers choose bright colors to use where immediate attention is required and light or soothing colors are used when there is no such requirement. Human pupil contracts while looking at bright colors and that causes muscular tiredness. The tiredness increases further if there is mixture of dull and bright colors, because pupil dilates on exposure to dull colors. This contraction and dilation of eye pupil reduces user's readability [4]. From this we conclude that such colors should be chosen which do not give muscular tiredness to users. *Bright colors should be used only where they are needed.*

### 2.9.3 Blue Colour as Background

The choice of background color is an important issue. If appropriate color is not chosen this may put cognitive load on the user and s/he may leave the system and switch to some other system. Blue color has short wavelength and fovea have few blue-sensitive cones which makes difficult for an eye to focus on it, and therefore making it an ideal background color [4] [10]. Hence from this we conclude that blue should be used as a background color, as it won't distract the user from the information presented on the screen.

## 2.10 Screen Layout

Layout of objects on the screen is a very important issue. Reading habits of humans can guide in this matter. It is important to understand how humans read. Human eye is first attracted to the most prominent and highly colored object among others present on the screen. Then it follows reading gravity. Reading gravity means eye moves down the screen and shifts from text to graphics and vice versa [4]. *Hence the most important object should be placed on the top half of the screen. Rest of the information follows it in the direction of reading gravity.*

## 2.11 Use of Graphics

Most of the interface designers use graphics in their design to lessen the visual monotony. But some times they neglect the fact that graphics used should be in accordance with the information that is displayed on the screen [4]. If this is the case, user spends more time in making relationship between the information and graphics. S/he will try to figure out why that graphic is used while there is no semantic linkage between the two. *Therefore graphics should be chosen according to the content presented on screen, while making sure that their purpose is not just to fill the space.*

## 2.12 Typography

Some people consider that text written in capitals is easier to read because it is bigger, but the facts show that the reverse is true. Text written in capital letters shows emphasis but reduces reading speed by 12% and uses up 30% more space than proportionally spaced characters [4]. Shape of characters become identical in the case of capitals as compared to the lower case characters. This fact is described in the following figure.



Fig. 6. Difference between small and capital letters [4]

*Hence capitals should be restricted to where more attention is required and should not be used in bulk of text.*

### 3 Conclusion

In this paper we presented some guidelines to design user interfaces. The guidelines will help designers in constructing more useful interfaces that would help a user in interacting with the system efficiently and with reduced cognitive load.

### References

1. Aberg G., and Chang J. Applying Cognitive Science Research In Graphical User Interface (GUI) <http://www.dh.umu.se/default.asp?-naar=2005&p=1692> (2005)
2. Passer M.W., Smith R.E.: Psychology Frontiers and applications, Mac GrawHill (2001)
3. Anderson J. R.: Cognitive Psychology and its implications, 6th edition, Worth Publishers (2005)
4. Wynn, S.: Design principles for slides and overheads. In Summers, L. (Ed), A Focus on Learning, p287-291. Proceedings of the 4th Annual Teaching Learning Forum, Edith Cowan University., Perth: Edith Cowan University (1995)
5. Withrow J.: Cognitive Psychology and Informaton Architecture: From Theory to Practice: [http://www.boxesandarrows.com/view/cognitive\\_psychology\\_ia\\_from\\_theory\\_to\\_practice](http://www.boxesandarrows.com/view/cognitive_psychology_ia_from_theory_to_practice) (2003)
6. Francis G., Neath I., Mackewn A., Goldthwaite D.: Cog Lab Student Manual for 36 Experiments, Wordsworth (2003)
7. Preece J., Rogers Y., Sharp H.: Interaction Design: Beyond Human Computer Interaction 2nd edition, John Wiley and Sons, Ltd (2002)
8. Shneiderman B.: Designing the User Interface, 2nd eidition, Addison Wesley (1997)
9. Dix A., Finlay J., Abowd G. D., Beale R.: Human-Computer Interaction, 3rd edition, Pearson Education Ltd (2004)
10. Cos K., Walker D.: User Interface Design”, 2nd edition, Prentice Hall (1993)
11. Solso, R. L., Maclin, O. H., Maclin M. K.: Cognitive Psychology 8th edition, Pearson <http://www.drippingsprings.com/> (2008)

### Appendix

#### Summary of Guidelines

1.	Text should be placed in reader's right visual field (Placement of text)
2.	Images should be placed in reader's left visual field (Image Placement)

3.	Relative location of stimulus and response should be same (Simon effect)
4.	Interface layout should be according to previous experience of user (Positive transference)
5.	Interface should be according to the mental images of user (Mental imagery)
6.	Information presented to user should be based upon previous knowledge of user (Transference)
7.	Number of steps to accomplish a task should be kept small (Memory)
8.	Icons on graphical user interface should convey their meaning to users (Perception)
9.	Sounds and audios should be audible and convey the message clearly (Perception)
10.	Text should be visible and its color should not conflict with background (Perception)
11.	Interface should provide only basic features to novice and advanced features to experienced users (Learning)
12.	Background music should be avoided along with other audio information (Limitation of motor system)
13.	Background image should be avoided especially with text on it (Limitation of motor system)
14.	An interface should not contain more than $7 \pm 2$ chunks of information (Capacity of short term memory)
15.	Unimportant information should not be presented in a way to interfere user (Minimize interference)
16.	Use minimum number of colors ( $7 \pm 2$ ) on screen at a time (Colors)
17.	Avoid many bright colors (Colors)
18.	Use blue as background color (Colors)
19.	Place most important information in the top half of screen (Screen layout)
20.	Use of graphics should be in accordance with the information presented on the screen (Use of graphics)
21.	Capital letters should be used to give emphasis only and not in bulk of text (Typography)
22.	Information should be categorized to facilitate visual search (Visual search)

23.	Category label should fully depict the main theme of information (Visual search)
24.	Category label should serve as hints for search (Visual search)
25.	To avoid readers making wrong relationships rules of proximity and similarity should be kept in mind

# CAST: A Novel Trajectory Clustering and Visualization Tool for Spatio-Temporal Data

Hazarath Munaga, Lucio Ieronutti and Luca Chittaro

HCI Lab, DIMI, University of Udine, Udine, Italy.  
hazarath.munaga@gmail.com,{lucio.ironutti, luca.chittaro}@dimi.uniud.it

**Abstract.** This paper presents a novel technique for clustering and visualizing spatio-temporal data to analyze the navigational behavior of moving entities, such as users, virtual characters or vehicles. For testing our proposal, we developed CAST (Clustering And visualization tool for Spatio-Temporal data), a tool designed for interactively studying moving entities navigating through real as well as virtual environments. Such analysis allows one to derive information at a level of abstraction suitable to support (i) the evaluation of user spaces and (ii) the identification of the predominant navigation behavior of users. We demonstrate the effectiveness of our solution by testing the tool on data acquired by recording the movements of users while navigating through a virtual environment.

## 1 Introduction

Widespread use of sensor networks and location aware devices has resulted in large amounts of spatio-temporal datasets in a variety of different contexts. The number and size of these datasets continues to increase rapidly, making their manual analysis impossible. The analysis of moving entities is a crucial task in several application domains and different techniques have been proposed in the literature referring to data acquired both from real and virtual environments (hereinafter, VEs). In the literature there are two main categories of solutions: (i) *trajectory clustering techniques* to identify and classify relevant subsets (e.g., recurrent or unique navigational behaviors) from very large data repositories, and (ii) *visual techniques* targeted at supporting the analysis by displaying trajectories in an effective (and intuitive to interpret) way.

In [1] is presented a variation of *expectation maximization algorithm* [2] to cluster small sets of trajectories obtained from real environment. However, their method is a model-based approach and therefore characterized by scalability problems. A novel method based on *non-parametric statistics* for clustering time-series data has been proposed in [3]; authors tested their proposal on the dataset of mouse mammary gland development, showing that their solution is able (i) to match a manual clustering by a domain expert and (ii) to cluster groups of genes with known related functions. However, the proposed solution is based on number

of time points, and as the number of time points increases, the number of trajectories increases exponentially and some cluster combining is required. Other authors (e.g.,[4][5]) use *artificial neural networks* for clustering trajectories. However, these solutions suffer from typical limitations of neural networks, such as hidden node complexity and the difficulty of modifying the network once it has been trained.

The second category of techniques for studying spatio-temporal data is based on visual analysis that ranges from approaches targeted at supporting the analysis of user interactions during the navigation on the Web, to solutions focused at highlighting navigation problems into VEs. Some of them [6][7] support both *non-aggregated* and *aggregated visualizations* of users navigation; *non-aggregated visualizations* separately display the path followed by each moving entity, while *aggregated visualizations* employ various techniques for displaying the navigational behavior of groups of users. The latter solution is particularly effective when the analysis concerns a huge number of different trajectories.

By integrating clustering and visualization techniques [8] proposed *ViEWNet*, a tool for detecting and visualizing the hierarchies of dense areas on spatial network by monitoring moving objects through the GPS receiver. They used color coded line and arrow drawings for visualizing the objects. However, the tool is based on *hierarchical partitioning/clustering*, and then it is sensitive to noise and outliers as well as difficulty in handling different shapes (e.g., convex shape) and sized clusters [9].

The previously discussed solutions can not answer queries like “Which individuals of a population move together?” or “Find groups of entities that perform similar sequences of changes or non-changes in their direction?”. In this paper we present CAST (Clustering And visualization tool for Spatio-Temporal data), a tool designed for providing solutions to the above queries. The main goal of our research is to develop an effective solution to understand and explore the data by integrating visualization and clustering techniques.

## 2 Trajectory Clustering

Any clustering algorithm requires a dissimilarity method for calculating dissimilarity between trajectories. For example, [10] proposed the usage of Euclidean distance between time series of equal length as the measure of similarity and it has been generalized in [11] for subsequence matching. Here we used [12] for calculating dissimilarity between trajectories since it is suitable for increasing dimensions and mainly designed for calculating dissimilarity between trajectories.

Trajectories are grouped into clusters using a threshold that can be selected based on observations and required number of clusters. To support the analysis of the navigational behavior of users, the cluster routine provides (i) a group of clusters, each one characterized by unique navigation behavior, and (ii) a set of trajectories equal to the number of clusters, which is used for representing the

behavior of each cluster (hereinafter, *representative trajectory*). The cluster routine contains the following stages: (i) dissimilarity matrix for trajectories is computed using dissimilarity algorithm [12], (ii) trajectories are grouped into initial clusters using *Initialization Algorithm* (Fig 1), (iii) representative trajectories are computed for each cluster using *RepTraj Algorithm* (Fig 1), and (iv) recompute the clusters and their representative trajectories using *Re-cluster Algorithm* (Fig 1).

CAST supports the visual analysis by providing both non-aggregated and aggregated interactive visualizations. To get overall idea of clusters, the tool codes the number of trajectories contained in each cluster with a color; such solution allows the analyst to easily identify both outliers and recurrent navigational behaviors. Different mechanisms to convert numerical values into colors have been proposed in the literature [13]. CAST supports different color code mechanisms, but in this paper we use hue interpolation, where the blue and red color indicates respectively the outlier and recurrent nature of the cluster. Moreover, the tool adopts colored hedgehog arrows to represent the direction of movement, and the gap between two arrows provides information on moving speed of users.

#### *Initialization Algorithm*

- a. Take first sample as first cluster. Classify all the remaining trajectories into this cluster if they are within the *threshold*.
- b. Take a trajectory (sequentially) which is not already classified into any of the cluster and consider it as a new cluster. Take all the other trajectories which are not kept in any of the clusters and keep in this cluster if they satisfy the *threshold* limit.
- c. Repeat step b till no new cluster is added.

#### *RepTraj Algorithm*

For each Trajectory of cluster  $C$  calculate cumulative dissimilarity with all other trajectories of the same cluster  $C$ . Select the trajectory which is having minimum cumulative dissimilarity and take this as representative trajectory of that cluster.

#### *Re-cluster Algorithm*

- a. For each Trajectory calculate dissimilarity with all the  $K$  representative trajectories and classify to the cluster for which dissimilarity is low (if it is less than *threshold*).

Re-calculate representative trajectories using *RepTraj Algorithm*. Repeat above steps till there is no change in representative clusters of  $i^{\text{th}}$  iteration and  $i+1^{\text{th}}$  iteration.

**Fig. 1.** Algorithm for trajectory clustering

### 3 Using CAST in Practice

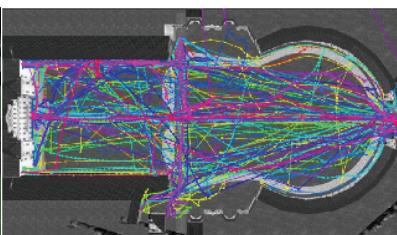
This section presents a detailed case study where CAST has been effectively used to analyze navigation logs of VillaManin3D [14], an online VE developed in VRML. VillaManin3D allows users to take a virtual walk into the Villa Manin Estate, located in the village of Passariano, Udine (Italy). As shown in Fig 2, the VE is characterized by two main parts separated by a central gate with side parapets of open wells. Users start their visit at the main entrance of the museum (bottom portion of Fig 2), and then they can freely navigate the VE. For example, they can start to visit the VE by navigating straightly through the open space towards the central gate of the museum, or moving through left and right corridors. To reach the main hall of the museum (upper portion of Fig 2), users can navigate through the central gate of the museum as well as the open space and corridors of the rectangular area.

For testing the tool, we remotely collected navigation data of 60 users containing 23293 samples, using a script on the client side sending at constant time intervals to our server the current position and orientation of the user. We started the analysis of collected samples by studying non-aggregated visualization of raw data (Fig 3(a)). This visualization allowed us to study in detail the behavior of each individual user. However, the non-aggregated visualization is not very effective if the analysis is targeted at extracting information on the general navigational behavior of users; in this case, it is more appropriate to use the aggregated visualizations based on proposed clustering technique.

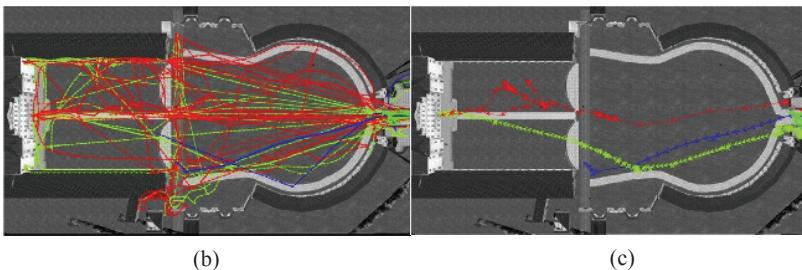
Fig 3(b) shows the obtained non-aggregated visualizations, where the above 60 trajectories are grouped into three clusters, and representative trajectories of the clusters are highlighted by increasing their line thickness. Fig 3(c) shows the obtained aggregated visualizations and direction of movement of the clusters. From Fig 3(c), we can observe that, after entering the museum, maximum users navigated through the middle portion of the museum, moderate users navigated through left side of the museum, and very less users navigated up to central gate of the VE. Moreover, it is interesting to note that initially users navigated very slowly at the entrance of the museum, but, once familiarized with the controls, users increased their navigation speed.



**Fig. 2.** Overview of the museum



**Fig. 3 (a)** Before clustering the trajectories



**Fig. 3.** Non-aggregated and Aggregated visualizations from CAST (a) before clustering the trajectories, obtained (b) clusters, and (c) representative trajectories of clusters

## 4 Conclusion and Future Work

In this paper, we presented a novel trajectory based clustering technique and ready-to-use tool for clustering historical spatio-temporal data. To show the effectiveness of the tool, we demonstrated it on a variety of trajectories moving with different directions and accelerations. Although we have recently started using CAST on real cases, results are very promising. The tool proved to be effective in supporting the analyst in discovering navigation patterns, identifying critical situations, and prompting usability improvements. For example, this tool can be used by Web3D designers to evaluate and improve the navigability of the VE, optimizing the user spaces and subsequently to improve the company return of interest.

There are lots of things to be done. First, based on the data acquisition and location technology, data can contain measurement errors and noise. For example, in the case of GPS technology, data accuracy is affected mainly by two factors. The first is the intrinsic measurement error of the GPS receiver, and the second is related to the sampling rate, which involves trajectory reconstruction process that approximates the movement of the entities between two localization points. More specifically, collected data are not guaranteed to be the outcome of sampling at fixed time intervals and sensors collecting the data may fail for some period of time leading to inconsistent sampling rates. For limiting type of errors, we plan to introduce in our tool a stage for removing or at least reducing errors and noise.

Second, generally complete trajectory cannot be similar with other trajectories but some portion of the trajectory will be similar, and our key observation is that analyzing complete trajectory could miss some similar portions of trajectories. Instead, if we extract similar portions of trajectories and analyze them we may get some interesting patterns. For this purpose, we intend to introduce in our tool a stage to segment trajectories using some criteria e.g., whenever the user changes his direction.

Finally, since the current version of the tool works for 2D space plus time, we intend to extent it for 3D space from both clustering and visualization point of view.

### Acknowledgements

Author, Hazarath Munaga would like to thank the Ministry of Italian Universities and Research (MIUR), Italy, for the financial support during his research work.

### References

1. Gaffney S., Smyth P.: Trajectory Clustering with Mixtures of Regression Models, in fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA (1999)
2. Dempster A. P., Laird N. M., Rubin D. B.: Maximum likelihood from Incomplete Data via the EM algorithm, *J. Royal Statistical Society, Series B.* 34, 1–38 (1977)
3. Phang T. L., Neville M. C., Rudolph M., Hunter L.: Trajectory Clustering: A Non-Parametric Method for Grouping Gene Expression Time Courses, with Applications to Mammary Development, in Pacific symposium on Bio computing (2003)
4. Cornia S., O'Hare G., Reily R.: Virtual Environment Trajectory Analysis: A Basis for Navigational Assistance and Scene Adaptivity *J. Future Generation Computer Systems* 21, 1157–1166 (2005)
5. Cornia S., Machine A.: Learning-Based Approach for Exploring User Spatial Mental Models, Kluwer Academic Publishers (2004)
6. Chittaro L., Ranon R., Ieronutti L.: VU-Flow: A Visualization Tool for Analyzing Navigation in Virtual Environments *J. Visualization and Computer Graphics* 12, 1475–1485 (2006)
7. Gopal P., Agata O., Yves J., Ingrid C.: Visualization of Sports Using Motion Trajectories: Providing Insights into Performance, Style, and Strategy, in VIS '01: Conference on Visualization '01, IEEE Computer Society, Washington, DC, USA 75–82 (2001)
8. Kriegel H. P., Peter K., Martin P., Matthias R.: ViEWNet: Visual Exploration of Region-Wide Traffic Networks,” in 22nd International Conference on Data Engineering, IEEE Computer Society, Washington, DC, USA pp. 166–166, (2006)
9. Tan P. N., Steinbach M., Kumar V.: Introduction to Data Mining, Addison-Wesley, (2006)
10. Agrawal R., Faloutsos C., Swami A.: Efficient similarity search in sequence databases,” in FODO '93, Fourth International Conference on Foundations of Data Organization and Algorithms, Springer-Verlag, London (1993)
11. Faloutsos C., Ranganathan M., and. Manolopoulos Y.: Fast subsequence matching in time-series databases, In ACM SIGMOD'94 Conference on Management of Data, ACM New York, USA (1994)
12. Laurinen P., Siirtola P., Roning J. Efficient algorithm for calculating similarity between trajectories containing an increasing dimension. In 24th IASTED International Conference on Artificial Intelligence and Applications, ACTA Press Anaheim, CA, USA (2006)

13. Schulze W. P., Tominski C., Schumann H.: Enhancing Visual Exploration by Appropriate Color Coding. In WSCG-05, International Conference Central Europe on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic (2005)
14. Villa Manin 3D Web site, Available at <http://udine3d.uniud.it>

# **Vote Stuffing Control in IPTV-based Recommender Systems**

Rajen Bhatt

Samsung India Software R&D Center, Logix Infotech Park, D-5, Sector-59,  
Noida-201301, India  
Email: {rajen.bhatt@samsung.com, rajen.bhatt@gmail.com}

**Abstract.** Vote stuffing is a general problem in the functioning of the content rating-based recommender systems. Currently IPTV viewers browse various contents based on the program ratings. In this paper, we propose a fuzzy clustering-based approach to remove the effects of vote stuffing and consider only the genuine ratings for the programs over multiple genres. The approach requires only one authentic rating, which is generally available from recommendation system administrators or program broadcasters. The entire process is automated using fuzzy  $c$ -means clustering. Computational experiments performed over one real-world program rating database shows that the proposed approach is very efficient for controlling vote stuffing.

## **1 Introduction**

Recommender systems automatically generate the list of recommended programs from the large number of programs offered by various channels / content distributors. This is generally done based on user profiles and/or in response to specific queries supplied by users [1, 2, 3]. In one way it can be seen as a content-based program retrieval system, where content may be extracted automatically or supplied by users as reviews (or ratings). In other way it can also be seen as an engine which mines existing program databases and newly added programs in the databases with reference to user profiles and recommends programs of interests to users.

Typical program recommender systems use viewers' ratings over a set of predefined program genres as back ground information along with algorithms to generate the recommendations. There have been many advances in recommender systems research. An extensive review of various approaches used in recommender systems and areas of improvements for current recommender systems have been presented by Burke [1] and Adomavicius and Tuzhilin [4], respectively.

The work presented in this research is towards a better method for representing user behavior so

that a less intrusive recommendation system can be built based on user ratings or voting of programs over multiple genres. While various IPTV viewers enter program ratings, there is a possibility of various attempts of vote stuffing by individuals more interested in changing the current or actual ratings of programs than giving their true opinion about it. Internet movie database, IMDb [5], to a certain extent does this by applying various filters on overall ratings. However, IMDB does this only for the final rating of movies and not over genres. The approach proposed here applies to genre ratings first and then process the overall rating based on the clustering logic.

This paper is organized as follows. In Section 2, we introduce notations and in parallel to that explain the proposed approach of controlling vote stuffing using fuzzy  $c$ -means algorithm. We also explain the processing of overall rating based on the clustering logic. Computational experiments have been presented in Section 3. Section 4 concludes the paper with directions for further research.

## 2 The Proposed Content Popularity Modeling Approach

Let, a set of  $n$  programs  $\{M_1, M_2, \dots, M_n\}$  be described by a set of  $p$  genres  $\{g_1, g_2, \dots, g_p\}$ . Each genre  $g_j; j=1, 2, \dots, p$  describe content of program like *Drama*, *Comedy* (or *Laughter*), *Epic*, *Suspense* etc...and restricted to lie in the range [0 5], i.e.,  $0 \leq g_j \leq 5, \forall j = 1, \dots, p$ . Let, each program  $M_i$  is reviewed by  $N_i$  IPTV viewers over  $p$  genres. The problem is to identify the viewers interested in *vote stuffing* only, instead of giving their true opinion over program content. Once this is done, overall program rating (i.e., *Outstanding*, *Average*, *Disappointing*) can be considered only for genuine ratings and rest all the ratings can be neglected by considering outliers or noise.

We handle this problem using the fuzzy  $c$ -means clustering initiated with a genuine center vector. Let, for a given movie  $M_i$ , review matrix  $\mathbf{R}_i$  is of dimension  $N_i \times p$  and  $r^{\text{th}}$  row is represented by the vector  $\mathbf{g}^{ir} = [g_1^{ir}, g_2^{ir}, \dots, g_p^{ir}]$ ;  $r = 1, \dots, N_i$ . We cluster  $\mathbf{R}_i$  into three fuzzy clusters using Fuzzy  $c$ -means algorithm [6], where one cluster represent genuine reviews while other two clusters represent extreme cases of *vote stuffing*.

It is usual practice to initiate the fuzzy  $c$ -means clustering program with the random partition matrix such that the sum of degrees of membership of an arbitrary pattern to all the fuzzy clusters is equal to one, i.e.,

$$\sum_{q=1}^k u_{iq} = 1, \forall i = 1, \dots, n, \quad (1)$$

where  $k$  is total number of clusters and  $u_{iq}$  is the degree of membership of  $i^{\text{th}}$  pattern to  $q^{\text{th}}$  cluster. Here we suggest initiating one fuzzy cluster center with the genuine review vector obtained from recommendation system administrators or program broadcasters. This review vector is only available for the program genres and not for the overall program rating. Let, the genuine review vector is represented by  $\mathbf{t}^i = [t_1^i, t_2^i, \dots, t_p^i]; i = 1, \dots, n$ . Initialize the first cluster center as  $\mathbf{t}^r$ ,  
*i.e.,*  $\mathbf{C}^{1i} = \mathbf{t}^i$ . (2)

Initialize rest of the two centers  $\mathbf{C}^{2i}$  and  $\mathbf{C}^{3i}$  as follows:

$$\begin{bmatrix} \mathbf{C}^{2i} \\ \mathbf{C}^{3i} \end{bmatrix} = \mathbf{R}_{i\_min} + \mathbf{RND} * (\mathbf{R}_{i\_max} - \mathbf{R}_{i\_min})$$

$\mathbf{R}_{i\_min}$  and  $\mathbf{R}_{i\_max}$  are  $1 \times p$  dimension vectors of minimum and maximum values of ratings over  $p$  genres, respectively and  $\mathbf{RND}$  is  $2 \times p$  dimension matrix of random values in the range  $[0 \ 1]$  taken from uniform distribution. The initial center matrix of dimension  $3 \times p$  is thus denoted as  
 $\mathbf{C}^i = [\mathbf{C}^{1i^T} \quad \mathbf{C}^{2i^T} \quad \mathbf{C}^{3i^T}]^T$ . With this initialization of centers perform the following iterations of fuzzy c-means clustering:

**Table 1.** Algorithm for fuzzy c-means with centers initialized

**Input:**  $m$  – exponent for fuzzyfication,  $\delta$  – termination criteria,  $\mathbf{C}$  – initial center matrix, and  $\mathbf{R}_i$

**Output:**  $\mathbf{U}$  – fuzzy partition matrix,  $\mathbf{C}$  – updated center matrix

```

check = 1
i=1
while (check)
    Calculate Euclidean distance matrix    $\mathbf{D} = d(\mathbf{C}, \mathbf{R}_i^r)$ 
    Calculate membership degree matrix  $\mathbf{U} = \frac{\mathbf{D}^{(\frac{1}{m}-1)}}{\sum \mathbf{D}}$ 
    Calculate new center matrix    $\mathbf{C} = \frac{\mathbf{U}^m * \mathbf{R}_i}{\sum \mathbf{U}^m}$ 
    Calculate Objective function  $Obj(i) = \sum \sum \mathbf{D}^2 * \mathbf{U}^m$ 
    if ( $Obj(i) - Obj(i-1) \leq \delta$ )
        Check = 0;
    end
end

```

Three fuzzy clusters for  $\mathbf{R}_i$  are denoted as  $f_{1i}, f_{2i}$ , and  $f_{3i}$ . Degrees of membership of  $\mathbf{g}^{ir}$  into these fuzzy clusters are represented by  $\mu_{f_{1i}}(\mathbf{g}^{ir}), \mu_{f_{2i}}(\mathbf{g}^{ir})$ , and  $\mu_{f_{3i}}(\mathbf{g}^{ir})$ . *True* review (or ratings) cluster is the one whose center is nearest to the genuine review vector  $\mathbf{t}^i$ . *True* review cluster for program  $M_i$  is thus represented by  $f_{ti}$  and is given by

$$f_{ti} = \arg \min_{t \in \{1, 2, 3\}} \{d_{1t}, d_{2t}, d_{3t}\}, \quad (3)$$

where  $d_{1t}, d_{2t}$ , and  $d_{3t}$  are Euclidean distances of three cluster centers from genuine review vector, respectively. Fuzzy clusters other than  $f_{ti}$  are described here as an attempt to *vote stuffing*.  $p$ -element review vector for  $M_i$  is represented by the center vector of the fuzzy cluster  $f_{ti}$ .

In addition to description by  $p$  genres, each program  $M_i$  has been rated in one of the three categories *{Outstanding, Average, Disappointing}* by each reviewer. For  $i^{\text{th}}$  program, category assigned by  $r^{\text{th}}$  reviewer is denoted as  $\text{cat}^{ir}$ . It is straight forward to mention that  $\text{cat}^{ir} \in \{\text{Outstanding, Average, Disappointing}\}, i=1, \dots, n, r=1, \dots, N_i$ . We consider category ratings of only those users which fall in true review cluster  $f_{ti}$ . Let,  $N_{ti}$  is number of users falling in  $f_{ti}$  and,  $N_{Oti}, N_{Ati}$ , and  $N_{Dti}$  are number of users who have rated program  $M_i$  as *Outstanding*, *Average*, and *Disappointing*, respectively. Precisely,

$$\begin{aligned} N_{Oti} &= \|\text{cat}^{ir}\|, \text{cat}^{ir} = \{\text{Outstanding}\}, \\ N_{Ati} &= \|\text{cat}^{ir}\|, \text{cat}^{ir} = \{\text{Average}\}, \\ N_{Dti} &= \|\text{cat}^{ir}\|, \text{cat}^{ir} = \{\text{Disappointing}\}. \end{aligned} \quad (3)$$

The overall rating for program  $M_i$  is fuzzy number with degrees of membership in *Outstanding*, *Average*, and *Disappointing* are given by the 3-element row vector

$$\mathbf{v}^i = \left[ \frac{N_{Oti}}{N_{ti}} \quad \frac{N_{Ati}}{N_{ti}} \quad \frac{N_{Dti}}{N_{ti}} \right] = [v^{Oi} \quad v^{Ai} \quad v^{Di}] \quad (4)$$

Here  $v^{Oi}$ ,  $v^{Ai}$ , and  $v^{Di}$  are referred as the degrees with which  $i^{\text{th}}$  program has been categorized as *Outstanding*, *Average*, or *Disappointing*. This formulation of assigning fuzzy class for each program fulfills the fuzzy  $c$ -means algorithm constraint of sum of degrees of membership of a pattern to  $q$  arbitrary fuzzy clusters should be one [6], *i.e.*,

$$\sum_{q=1}^k u_{iq} = 1, \forall i = 1, \dots, n. \quad (5)$$

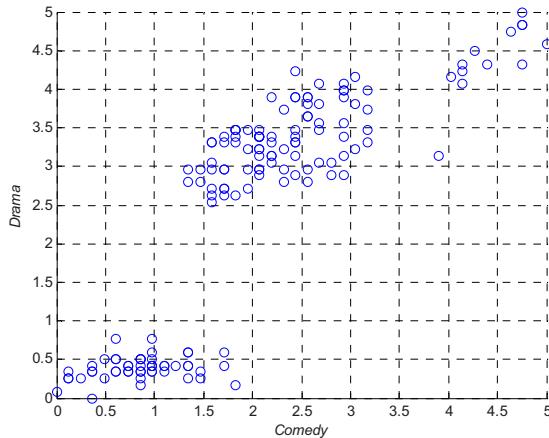
With this approach the overall program ratings are considered only for those peoples' reviews which fall in the *true* review fuzzy cluster. Further, the approach of modeling overall program review as a fuzzy set with degrees of memberships into *Outstanding*, *Average*, and *Disappointing* gives more insight to program

popularity instead of a single rating. Further, each IPTV user is allowed to rate the program only once. This information can easily be tracked by creating a program review log of IPTV users against programs rated.

In the next section, we present computational experiments with one real-world program review dataset.

### 3 Computational Experiments

We have considered here an arbitrary program reviewed by 150 users over two genres *comedy* and *drama*. Each review is constrained within the range of [0 5], where 0 is lowest and 5 is highest rating. The scatter plot of reviews is shown in Fig. 1.



**Fig. 1.** Scatter plot of reviews over *comedy* and *drama*

For the considered problem genuine review obtained from the recommender system administrator is  $[\text{comedy} \text{ drama}] = [2.5 \ 3]$ . We have initiated the clustering process with two random centers and one center  $[2.5 \ 3]$ . The initial centers to start the fuzzy  $c$ -means clustering are thus assigned as

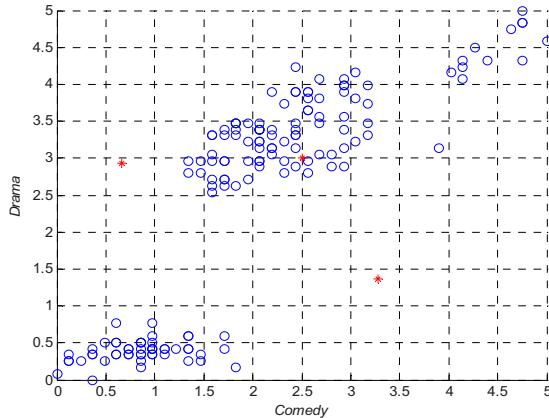
$$\begin{bmatrix} 2.50 & 3.00 \\ 3.28 & 1.37 \\ 0.66 & 2.93 \end{bmatrix}.$$

The initial cluster centers are shown in Fig. 2.

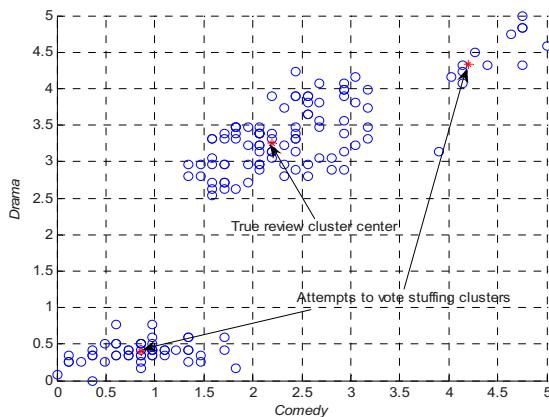
With this initial cluster centers we perform fuzzy  $c$ -means clustering outlined in Table I. Algorithm terminates after 24 iterations. Three cluster centers obtained after algorithm termination are shown in Fig. 3.

Three cluster centers obtained are

$$\begin{bmatrix} 2.18 & 3.25 \\ 4.19 & 4.32 \\ 0.85 & 0.40 \end{bmatrix}.$$



**Fig. 2.** Initial cluster centers shown by '\*' mark



**Fig. 3.** True review and vote stuffing cluster centers identified

Euclidean distances of these cluster centers with the genuine review [2.5 3] are {0.40, 2.15, 3.07}. Thus as per the proposed method cluster 1 with centers [2.18 3.25] is true review cluster  $f_{ti}$ . Number of users falling in  $f_{ti}$  are 85, i.e.,  $N_{ti} = 85$ . Out of 85, 57 reviewers have voted the program as *Outstanding*, 7 as *Average*, while 21 reviewers have voted the program as *Disappointing*. With this statistics, fuzzy number representing overall program rating is given by

$$\mathbf{v}^i = \left[ \begin{array}{ccc} 57 & 7 & 21 \\ 85 & 85 & 85 \end{array} \right] = [0.67 \quad 0.08 \quad 0.24]$$

In addition to this deeper insight into the overall program rating, the proposed methodology also identifies users (through IP addresses) who have similar tastes for several genres. For example, genuine review says that the *comedy* and *drama* content are average in the program. People belonging to cluster 1 are group of people who have similar tastes for *comedy* and *drama* genres. While, on the two extreme ends people belonging to cluster 2 are rating the existing contents as very high and people belonging to cluster 3 are rating the existing contents as very low for *comedy* and *drama*. Using this information certain clues can be derived on the various customized content delivery scheme for users falling into various categories.

The original review for the  $i^{\text{th}}$  program is thus represented by the center of the true review cluster and overall review as a fuzzy number spanning across outstanding, average, and disappointing.

## 4 Conclusion and Further Work

In this paper, we have presented an approach to control *vote stuffing* attempts using fuzzy  $c$ -means clustering algorithm. The approach has been validated over real-world program review dataset for 150 reviewers. As a further work, we will attempt on building a recommender system using collaborative filtering over true review clustered data. We will also explore the possibility of using one pass clustering instead of fuzzy  $c$ -means clustering. While deploying in real-time analysis of IPTV reviews scalability and fast execution time are some of the important aspects of algorithm. We will work towards the improvement in time and space complexity of fuzzy  $c$ -means clustering and possible one-pass clustering like Gaussian mixture models.

The initial prototype will be implemented and tested for IPTV platform where users are allowed to feed reviews for the program.

## References

1. Burke R.,: Hybrid recommender systems: survey and experiments, *User Modeling and User Adapted Interaction*. 12, 331–370 (2002)
2. Resnick, P. , Varian, H.R.: Recommender systems, *Communications of the ACM*, 40, 56–58 (1997)
3. Kautz, H.: *Recommender Systems*, AAAI Press, Menlo Park, CA, (1998)
4. Adomavicius, G., Tuzhilin, A. Towards the next generation of recommender systems: a survey of the state-of-the-art and possible extensions”, *IEEE Transactions on Knowledge and Data Engineering*, 17, 734–749, (2005)
5. The Internet Movie Database, <http://www.imdb.com/>
6. Bezdek J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. NY: Plenum Press, (1982)

# **Static and Dynamic Features for Improved HMM based Visual Speech Recognition**

R. Rajavel and P. S. Sathidevi

National Institute of Technology Calicut, India  
{rettyraja@gmail.com, sathi@nitc.ac.in}

**Abstract.** Visual speech recognition refers to the identification of utterances through the movements of lips, tongue, teeth, and other facial muscles of the speaker without using the acoustic signal. This work shows the relative benefits of both static and dynamic visual speech features for improved visual speech recognition. Two approaches for visual feature extraction have been considered: (1) an image transform based static feature approach in which Discrete Cosine Transform (DCT) is applied to each video frame and 6x6 triangle region coefficients are considered as features. Principal Component Analysis (PCA) is applied over all 60 features corresponding to the video frame to reduce the redundancy; the resultant 21 coefficients are taken as the static visual features. (2) Motion segmentation based dynamic feature approach in which the facial movements are segmented from the video file using motion history images (MHI). DCT is applied to the MHI and triangle region coefficients are taken as the dynamic visual features. Two types of experiments were done one with concatenated features and another with dimension reduced feature by using PCA to identify the utterances. The left-right continuous HMMs are used as visual speech classifier to classify nine MPEG-4 standard viseme consonants. The experimental result shows that the concatenated as well as dimension reduced features improve the visual speech recognition with a high accuracy of 92.45% and 92.15% respectively.

## **1 Introduction**

Many researchers were trying to design Automatic Speech Recognition (ASR) systems which can understand human speech and respond accordingly [10]. However, the performances of the past and current ASR systems are still far behind as compared to human's cognitive ability in perceiving and understanding speech [1]. The weakness of most modern ASR systems are their inability to cope robustly with audio corruption which can arise from various sources, for example environment noises such as engine noise or other people speaking, reverberation effects or transmission channel distortion etc. Thus one of the main challenges

being faced by the ASR research community is how to develop ASR systems which are more robust to these kinds of corruptions that are typically encountered in real-world situations. One approach to this problem is to introduce another modality to complement the acoustic speech information which will be invariant to these sources of corruption [1]. Visual speech is one such source, obviously not perturbed by the acoustic environment and noise. Such systems that combine the audio and visual modalities to identify the utterances are known as audio-visual speech recognition (AVSR) system [1]. The first AVSR system was reported in 1984 by Petajan [11] since then, over a hundred articles have concentrated on AVSR with the vast majority appearing during the last decade [3]. AVSR systems can enhance the performance of the conventional ASR systems especially under noisy condition [1]. This work describes the part of AVSR system known as visual speech recognition (VSR) for viseme classification.

Visual features proposed in the literature of AVSR or VSR systems can be categorized into shape-based, pixel-based and motion-based features [9]. The first AVSR system [11] used shape-based features such as height, width, and area of the mouth derived from binary mouth images for isolated word recognition. Since then, many approaches have been developed which includes active contours (often called snakes), deformable templates, active shape models etc. [3]. The VSR systems that use pixel-based features assume that pixel values around the mouth area contain salient speech information. As with shape-based feature approaches, there have also been numerous studies using different image transform methods. These methods include discrete cosine transform (DCT) [2,3,7,8], discrete wavelet transform (DWT) [2,3,7,8], principal component analysis (PCA) [3,7,8] and linear discriminant analysis (LDA)[3,7].

Pixel-based and shape-based features are extracted from static frames hence viewed as static features. Motion-based features are features that directly utilize the dynamics of speech [5]. Dynamic features are better in representing distinct facial movements and static features are better in representing oral cavity that cannot be captured either by lip contour or motion-based features, hence this work focus the relative benefits of both static and dynamic features for improved VSR.

Second section, describes details of the experimental database and image transform (DCT) based static feature extraction is explained in section three. The fourth section describes the motion segmentation based dynamic feature extraction. The fifth section tells the reason for feature concatenation and feature dimensionality reduction by using PCA and also the HMMs model parameters. The result of the experiment is reported and discussed in section six. The conclusion and future direction of this work are outlined in the last section.

## 2 Experimental Database

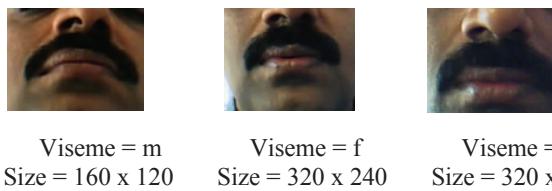
This work adopts MPEG-4 standard visemes for classification. 9 viseme consonants are chosen for the experiment and are highlighted in bold letter in Table 1. Potamianos et al. has demonstrated that using mouth videos captured

from cameras attached to wearable headsets produced better results as compared to full face videos [12]. With refer to the above, as well as to make the system more practical in real world environment, videos focusing speaker's mouth region were recorded using an inexpensive web camera in a typical office environment. Advantage of this kind of arrangement is that face detection, mouth location estimation and identification of the region of interest etc. are no longer required and thereby reducing the computational complexity.

Most of the audio-visual speech database available are recorded in ideal studio environment with controlled lighting or kept some of the factors like background, illumination, distance between the camera and speaker's mouth, view angle of the camera etc as constant. But in this work, the recording was done on different days with different values for the above factors to make the system more practical. The bold letter visemes in Table 1 were recorded 25 times and stored as true color (.AVI) files, samples of the recorded video illustrating the above variations are shown in Fig.1.

**Table 1.** Viseme model of the MPEG-4 standard for English consonants

Viseme No.	1	2	3	4	5	6	7	8	9
Viseme Categories	p, b, <b>m</b>	f, v	<b>th</b> , dh	t, <b>d</b>	k, g	sh, zh	s, <b>z</b>	n, <b>l</b>	<b>r</b>



Viseme = m  
Size = 160 x 120      Viseme = f  
Size = 320 x 240      Viseme = l  
Size = 320 x 240

**Fig. 1.** Sample videos illustrating different variations in size, view angle, background and Illumination etc.

### 3 Image Transform (DCT) based Static Feature Extraction

I.Matthews et al. [3] reported that intensity-based features using discrete cosine transform outperform model-based features. Hence DCT is employed in the proposed work to represent static features.

### 3.1 Preprocessing

Prior to the image transform, the recorded mouth region video represented as  $\{A_t(m,n) : 1 \leq t \leq 60\}$  are converted to equivalent RGB image. This RGB image is converted to YUV color space and only the Y channel is kept, as this retains the image data least affected by the video compression. The resultant Y channel was sub sampled to 16 X 16 and then passed as the input  $\{A_t(m,n) : 1 \leq m, n \leq 16\}$  to the DCT. The images of 16 X 16 pixels provided slightly better performance than 32 X 32 pixel images [2], and hence in this work 16 X 16 pixel images are taken as input to the DCT.

### 3.2 DCT Triangle Region Features

The DCT represents an image as a sum of sinusoids of varying magnitude and frequencies. It has the property that, most of the visually significant information about the image is concentrated in just a few coefficients of the DCT.

The two-dimensional DCT of an m-by-n image sequence  $\{A_t(m,n) : 1 \leq t \leq 60\}$  is defines as

$$B_t(p,q) = \left\{ \frac{1}{\sqrt{2N}} C(p) C(q) \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A_t(m,n) \cos\left[\frac{(2m+1)p\pi}{2M}\right] \cos\left[\frac{(2n+1)q\pi}{2N}\right] \quad 0 \leq p \leq M-1; 0 \leq q \leq N-1 \right\}$$

$$\text{Where, } M = N = 16; \quad \text{and} \quad C(x) = \begin{cases} 1/\sqrt{2} & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases}$$

The DCT return a 2D matrix  $B_t(p,q)$  of coefficients and more ever they are packed by ascending frequency in diagonal lines. Also proved that triangle region feature selection outperforms the square region feature selection, as those include more of the coefficients corresponding to low frequencies [2]. In some previous studies using DCT [8] including the DC component gave slightly improved recognition performance. This work also included the DC component and 6 X 6 triangle region coefficients were chosen as the static feature of that frame,

$$O_t = \{B_t(1,1) B_t(1,2) B_t(1,3) B_t(1,4) B_t(1,5) B_t(1,6) B_t(2,1) B_t(2,2) B_t(2,3) B_t(2,4) \\ B_t(2,5) B_t(3,1) B_t(3,2) B_t(3,3) B_t(3,4) B_t(4,1) B_t(4,2) B_t(4,3) B_t(5,1) B_t(5,2) \\ B_t(6,1)\} \quad : 1 \leq t \leq 60\}$$

The Principal component analysis is applied over all 60 features corresponding to video frames to reduce the inter frame redundancy, and the resultant 21 principal component coefficients are taken as the final static features for recognition.

## 4 Motion Segmentation based Dynamic Feature Extraction

In this work, dynamic visual speech features which show the facial movements of the speaker are segmented from the video data using an approach called motion history images (MHI) [5]. MHI is a gray scale image that shows where and when movements of speech articulators occur in the image sequence.

Let  $\{A_t(m,n) : 1 \leq t \leq 60\}$  be an image sequence, the difference of frame is defined as

$$DIF_t(m,n) = |A_t(m,n) - A_{t-1}(m,n)| \quad (1)$$

Where  $A_t(m,n)$  is the intensity of each pixel at location  $(m,n)$  in the  $t^{\text{th}}$  frame and  $DIF_t(m,n)$  is the difference of consecutive frames representing region of motion. Binarization of the difference image  $DIF_t(m,n)$  over a threshold  $\tau$  is

$$DOF_t(m,n) = \begin{cases} 1 & \text{if } DIF_t(m,n) \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The value of threshold  $\tau$  is optimized through experimentation. Finally MHI  $(m,n)$  is defined as

$$MHI(m,n) = \text{Max} \bigcup_{t=1}^{N-1} DOF_t(m,n) * t \quad (3)$$

Where  $N$  represents the number of frames used to capture the mouth region motion. In equation (3), to show the recent movements with brighter value, the binarised version of the DOF is multiplied with a ramp of time and integrated temporally. Next, DCT was applied to MHI  $(m,n)$  and the transform matrix  $D(m,n)$  is obtained. Similar to static feature extraction, only triangle region coefficients were considered as the dynamic features, they are,

$$\begin{aligned} O = \{ & D(1,1) D(1,2) D(1,3) D(1,4) D(1,5) D(1,6) D(2,1) D(2,2) D(2,3) D(2,4) \\ & D(2,5) D(3,1) D(3,2) D(3,3) D(3,4) D(4,1) D(4,2) D(4,3) D(5,1) D(5,2) \\ & D(6,1) \} \end{aligned}$$

## 5 The HMM based VSR System

### 5.1 Feature Concatenation and Dimensionality Reduction

The feature concatenations were already done by many researchers. The dynamic features were obtained by derivative of static features and concatenated [2]. The

shape-based feature was concatenated with appearance-based feature [3]. The key reason for feature concatenation is that static features are better in representing significant speech reading information lies within the oral cavity that cannot be captured by the dynamic features [8]. Similarly the dynamic features are better in representing distinct facial movement, it means that, complementary information exist in both feature sets, hence the feature concatenation will perform better than the individual features. In this work additionally the 42 dimensional concatenated features were reduced to 21 dimensional features by using PCA, and both of these features were used independently for HMM based viseme consonants classifications.

## 5.2 HMM Specifics

HMM is a finite state network based on stochastic process. The left-right HMM is a commonly used classifier in speech recognition, since it has the desirable property that it can readily model the time-varying speech signal [10]. This work adopts single-stream, continuous HMM to classify the concatenated visual speech features. Each viseme in the database is modeled using a left-right HMM with 12 states and 6 Gaussian mixtures per state. During training of the HMM, the transition probability and the observation probabilities are estimated iteratively using Baum-Welch training algorithm. In the classification stage, the unknown features are presented to the trained HMMs and the features are assigned to the viseme class whose HMM produces output with the highest likelihood [6].

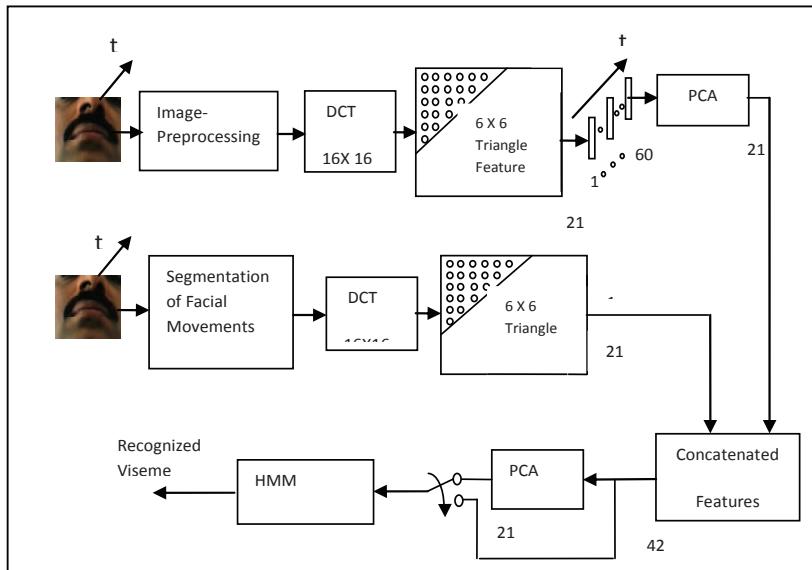
## 6 Experimental Results and Discussion

This work concentrates on classification of MPEG-4 standard viseme consonants only, not vowels, because vowels are easily distinguishable either by using static or dynamic features alone [4, 6, and 9]. For each consonant viseme class, two sets of features were extracted, one representing static visual information and other representing dynamic visual information. In static feature extraction, each video file containing speaker's mouth region was converted to YUV color space and sub sampled to 16 X 16 pixels image sequences. Next DCT was applied to each image and subsequently PCA was used to reduce the redundancy of inter frame features, the resultant 21 coefficients were taken as the static features. In dynamic feature extraction, from each video file one MHI was generated, and subsequently DCT was applied to obtain the dynamic features with the dimension of similar 21 coefficients.

Next, the static and dynamic features were concatenated and given as the input to the HMM classifier. Since both of these features are DCT coefficients there may have some redundancy which was reduced further by PCA, and the resultant reduced 21 dimensional features was also given to the classifier. The proposed visual speech recognition system is shown in Fig.2. For each viseme class 20 video files were given for training the HMM and 5 video files were tested. This

process was repeated for both the feature sets independently 15 times with different sets of training and testing data. The average recognition rates of the HMMs for the 15 repetitions were computed and tabulated in Table 2. The average recognition rate of the proposed VSR system using concatenated features is 92.45% and using dimension reduced features is 92.15%.

Same MPEG-4 standard viseme recognition experiment were done using dynamic features in [5] and [6], and obtained the recognition rates of 84.7% and 83.33% respectively. Similar experiment using shape-based features were done in [4] and [9], and for the consonants, the recognition rate obtained by them were 84.7% and 84.6% respectively. Compared to the above experiments the proposed concatenated features improve the recognition accuracy to 92.45% and dimension reduced features to 92.15%. This improvement shows that complementary information exist in static and dynamic features of visual speech, for example with respect to concatenated features the viseme / d / and / r / had 65.3% and 70.6% recognition accuracy by using dynamic features alone, but by concatenating with static features, 92% and 90.6% of recognition accuracy were obtained.



**Fig. 2.** Proposed Visual Speech Recognition System

**Table 2(a).** Recognition accuracy of viseme consonants using only static features

Test No.	d	z	l	th	f	k	m	sh	r	Test Average (%)
1	5	5	5	4	5	5	5	4	4	93.33
2	5	5	5	5	5	5	5	3	5	95.56
3	5	3	5	5	4	5	5	4	5	91.11
4	3	4	4	5	4	5	4	4	5	84.44
5	5	5	5	5	5	5	5	5	5	100
6	5	5	5	4	3	5	5	3	3	84.44
7	5	5	4	4	5	5	5	5	4	93.33
8	5	5	4	5	4	5	5	4	5	93.33
9	3	5	4	5	5	4	4	5	5	88.89
10	5	5	4	4	5	5	5	5	4	93.33
11	5	4	5	5	5	5	5	5	5	97.78
12	4	4	5	5	5	4	5	1	3	80
13	5	4	4	5	4	5	5	5	5	93.33
14	5	5	4	4	5	5	5	5	3	91.11
15	3	5	4	5	4	5	5	2	5	84.44
Viseme Average(%)	90.6	92	89.3	93.3	90.6	97.3	97.3	80	85.3	90.65

**Table 2(b).** Recognition accuracy of viseme consonants using only dynamic features

Test No.	d	z	l	th	f	k	m	sh	r	Test Average (%)
1	4	3	4	4	5	4	5	4	3	80
2	4	4	3	3	3	4	4	5	4	75.56
3	3	4	2	5	5	5	3	4	3	75.56
4	4	3	4	5	4	1	5	5	4	77.78
5	4	5	2	5	1	5	5	5	5	80
6	5	4	5	5	5	4	4	5	2	86.67
7	4	4	5	5	5	3	5	5	4	88.89
8	2	2	3	4	5	3	5	5	5	75.56
9	3	5	4	4	4	3	4	5	3	77.78
10	3	3	4	5	4	5	1	5	4	75.56
11	2	4	5	4	5	4	5	4	4	82.22
12	3	5	2	5	4	3	5	5	2	75.56
13	4	4	5	5	4	4	5	5	3	86.67
14	1	4	2	5	5	4	5	5	4	77.78
15	3	4	5	5	4	4	5	5	3	84.44
Viseme Average(%)	65.3	77.3	73.3	92	84	74.6	88	94.6	70.6	80.02

**Table 2(c).** Recognition accuracy of viseme consonants using concatenated features  
(42 coefficients)

Test No.	d	z	l	th	f	k	m	sh	r	Test Average (%)
1	4	5	4	5	5	5	5	5	5	95.56
2	4	4	5	5	5	4	4	5	5	91.11
3	5	3	4	5	5	5	4	5	4	88.89
4	4	5	5	4	5	5	5	5	5	95.56
5	5	5	4	5	5	5	5	5	5	97.78
6	5	5	4	3	4	5	5	5	2	84.44
7	5	5	5	4	5	5	5	5	5	97.78
8	5	5	5	5	5	5	2	5	5	93.33
9	4	5	4	3	4	5	5	5	4	86.67
10	5	4	5	3	5	5	5	5	5	93.33
11	5	4	5	5	5	5	5	5	4	95.56
12	3	4	3	5	5	4	2	5	5	80
13	5	4	5	5	5	5	4	5	5	95.56
14	5	4	5	5	5	5	5	5	5	97.78
15	5	5	4	5	5	5	5	4	4	93.33
Viseme Average (%)	92	89.3	89.3	89.3	97.3	97.3	88	98.6	90.6	92.45

**Table 2(d).** Recognition accuracy of viseme consonants using dimension reduced features (21 coefficients)

Test No.	d	z	l	th	f	k	m	sh	r	Test Average (%)
1	3	5	5	4	5	4	5	4	4	86.67
2	3	5	5	3	5	5	4	4	3	82.22
3	5	5	5	5	5	5	5	4	5	97.78
4	5	5	5	4	4	5	5	4	4	91.11
5	4	5	5	5	5	5	5	5	5	97.78
6	5	4	5	5	5	5	5	5	5	97.78
7	5	4	5	5	5	5	5	5	3	93.33
8	5	3	4	4	4	5	5	4	5	86.67
9	4	4	4	5	4	5	5	5	5	91.11
10	5	5	5	5	5	5	5	5	5	100
11	5	5	4	5	4	5	5	3	4	88.89
12	4	4	5	5	5	5	5	5	5	95.56
13	4	5	4	5	5	5	4	4	4	88.89
14	5	4	5	5	5	5	5	5	4	95.56
15	4	5	5	5	4	4	4	4	5	88.89
Viseme Average (%)	88	90.6	94.6	93.3	93.3	97.3	96	88	88	92.15

Similarly the viseme / sh / had 80% recognition accuracy by using static features alone, but by concatenating with dynamic features, 98.6% of recognition accuracy were obtained. Similarly with respect to dimension reduced features, the visemes /l/, /f/, /sh/, and /r/ had 89.3%, 90.6%, 80% and 85.33% recognition accuracy respectively by using static features alone, but by including the dynamic features and reduce the dimension to 21 coefficients by PCA, the accuracy were improved to 94.5%, 93.3%, 88% and 88% respectively.

## 7 Conclusion

This work reports the design of a practical voiceless speech recognition system. The experimental data was recorded in a real office environment and recognized with the accuracy of 92.45% by using concatenated features with 42 coefficients and 92.15% by using dimension reduced features with only 21 coefficients. The

experimental result shows that concatenated features as well as dimension reduced features improves the performance of VSR and suitable to use in noisy environment and for voiceless communication in office environment. In future, the authors intend to test the system on a large vocabulary database covering words in local language and to test the performance with various video corruptions.

## References

1. Chen, T.: Audio Visual Speech Processing: Lip Reading and Lip Synchronization. *IEEE Signal Processing Magazine*, pp. 9–21 (2001)
2. Rowan S., Darryl S., Ji Ming M.: Comparison of Image Transform-Based Features for Visual Speech Recognition in Clean and Corrupted Videos. *EURASIP Journal on Image and Video Processing*. Hindawi Publishing Corporation (2008)
3. Potamianos, G., Neti, C., Luettin, J., Matthews, I.,: Audio-Visual Automatic Speech Recognition:An Overview. In: *Issues in Visual and Audio-Visual Speech processing*. Bailly, G., Vatikiotis-Bateson, E., Perrier, P. (eds.) MIT Press (2004)
4. Say Wei Foo, Yong Lian, Liang Dong: Recognition of Visual Speech Elements Using Adaptively Boosted Hidden Markov Models. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 14, No. 5 (2004)
5. Yau, W.C., Kumar, D.K., Arjunan, S.P.,: Voiceless Speech Recognition Using Dynamic Visual Speech Features. *Australian Computer Society, Inc*, pp. 93–101. Canberra Australia (2006)
6. Yau, W.C., Kumar, D.K., Weghorn, H.,: Visual Speech Recognition Using Motion Features and Hidden Markov Models. In: Kampel, M., Hanbury, A. (eds.) LNCS, pp. 832–839. Springer, Heidelberg (2007)
7. Potamianos, G., Verma, A., Neti, C., Iyengar, G., Basu, S.,: A Cascade Image Transform for Speaker Independent Automatic Speech Reading. In: *Proceeding of IEEE International Conference on Multimedia and Expo*. pp. 1097–1100, New York (2000)
8. Potamianos, G., Graf, H.P., Cosatto, E.,: An Image Transform Approach for HMM Based Automatic Lip Reading. In: *Proc of the International Conference on Image Processing*. Vol. 3, pp. 173–177, Chicago (1998)
9. Foo, S.W., Dong, L.,: Recognition of Visual Speech Elements Using Hidden Markov Models. In: Chen, Y.C., Chang, L.W., Hsu, C.T. (eds.) LNCS, pp. 607–614. Springer, Heidelberg (2002)
10. Rabiner, L.R., Juang, B.H.,: Fundamentals of Speech Recognition. *Signal Processing Series*, Prentice-hall, Englewood Cliffs, NJ (1993)
11. Petajan, E.D., Bischoff, B., Bodoff, D.,: An Improved Automatic Lip Reading System to Enhance Speech Recognition. In: *ACM SIGCHI-88*. pp. 19–25 (1988)
12. Potamianos, G., Neti, C., Huang, J., Connell, J.H., Chu, S., Libal, V., Marcheret, E., Hass, N., Jiang, J.,: Towards Practical Development of Audio-Visual Speech Recognition. In: *IEEE International Conference on Acoustic, Speech, and Signal Processing* (2004)

# A Single Accelerometer based Wireless Embedded System for Predefined Dynamic Gesture Recognition

Rahul Parsani and Karandeep Singh

Department of Electrical, Electronics and Instrumentation  
Birla Institute of Technology and Science - Pilani, Goa Campus, Off NH 17B,  
403726 Goa, India

{f2005316, f2006193}@bits-goa.ac.in

**Abstract.** The use of hand gestures provides an attractive alternative to cumbersome interface devices for human-computer interaction. A complete embedded system which facilitates the data acquisition, analysis, recognition, and the transmission wirelessly, of human dynamic gestures to a computer, is described. An intuitive algorithm for processing the accelerometer data was implemented and tested. This method permits all the analysis to be done by the embedded system processor. The system is capable of recognizing gestures involving a combination of straight line motions in three dimensions. These gestures are then used to control a host computer which executes tasks based on the gesture received. A sample application showing how the gestures can be mapped to the English alphabet is also shown.

## 1 Introduction

In the field of Human Computer Interaction (HCI), gesture recognition is becoming increasingly important as an interface method [1]. In particular, interpretation of human hand and body gestures can help in achieving the ease and comfort desired for HCI. In this paper we propose an accelerometer based wireless embedded system that can accurately detect predefined dynamic gestures. We consider a gesture as a stochastic process which exists at multiple levels in a hierarchy — simple, discrete movements correspond to subgestures at the lowest level, and combinations of these movements as complete gestures at a higher level.

The embedded system is fitted with a single 3 – axis accelerometer and in its current version can precisely detect a sequence of straight line motion on different axes, i.e., “up”, “down”, “left”, “right”, “front” and “back”. These simple, discrete movements correspond to subgestures and can be combined together to form more complex motions. Complexity of a gesture is defined as the number of subgestures required to form the gesture. The analysis is handled entirely by the

system and the final results are transmitted wirelessly to the computer which executes a set of tasks based on the gesture received. Different tasks can be assigned for various applications. A set of gestures chosen wisely can provide quick access to common functions of a program.

The paper is organized as follows. Section 2 gives an overview of previous research in this area. Section 3 describes the System Design, both the hardware and the algorithms involved. Section 4 shows the experimental results achieved by the system. Finally Section 5 concludes the paper and discusses some of the future directions of the system.

## 2 Background

Both vision-based and accelerometer-based techniques have been proposed for gesture recognition in [4]. General weaknesses of optical tracking systems is that there must be line-of-sight between the camera and the tracked object [6] and cannot be used in much less constrained domains and are reliant on lighting conditions or camera calibration. Sensor based systems have the advantage that computationally less intensive calculations are involved.

Accelerometer based solutions are a promising alternative for gesture recognition in mobile environments. Gestures can be classified into static and dynamic. Static gestures are defined with respect to a fixed frame of reference while dynamic gestures have a moving frame of reference [8]. The Acceleration Sensing Glove (ASG) [7] is an example of static hand gestures based recognition on accelerometer information. This information is then used to control a mouse pointing device. Dynamic predefined human gesture recognition has been shown for sports video annotation [3] and in Micro-Input Device Systems [9].

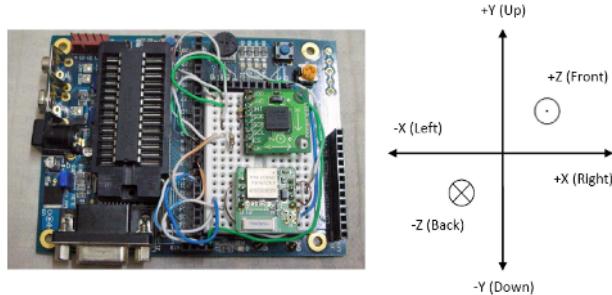
But, sensor information is inherently noisy. Algorithms for processing sensor data have been offered in [10] and [11]. Current research adopts techniques such as Hidden Markov Models (HMM) [13] [16], neural networks [2] or Finite State Machines (FSM) [14]. We have chosen to adopt an inertial measurement framework [12] which assumes prior knowledge of the gesture.

## 3 System Design

The system requirements are power efficiency, compactness and wireless connectivity and the components have been chosen to suit these constraints. The hardware consists of a microcontroller attached to the Bluetooth communication [15] and accelerometer sensor module. A picture of the system is shown in figure 1. The accelerometer is a STMicroelectronics LIS3LV02DQ 3-Axis - ±2g/±6g Digital Output Low Voltage Linear Accelerometer [17].

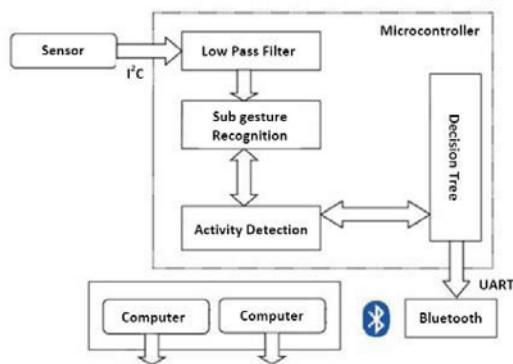
The accelerometer was connected to a Cypress Programmable System-on-Chip (PSoC) CY8C29466 [18] Mixed Signal Array Processor. The PSoC contains a Harvard architecture based processor with configurable peripheral blocks. It has

programmable pin configurations, clocking and dynamic reconfiguration which enable us to conserve even more power by powering up peripherals only when required during runtime. The PSoC uses a clock of 24MHz. The PSoC is connected to the accelerometer using the I<sup>2</sup>C communication protocol [5].



**Fig. 1.** Current state of the system and the reference axes for subgesture definition

The controller is connected to the wireless communication device, a Parani-ESD 210 Bluetooth module [19] through the UART protocol. The Bluetooth module has a communication range of 30 meters. We selected the Bluetooth protocol since it was readily available on most mobile computing devices and due to the simplicity involved in using Bluetooth Protocol stack. The block diagram of the system is below in figure 2.



**Fig. 2.** Block Diagram of the system.

The sampling rate of the system is 40Hz. The sampled data is then passed through a low pass filter to remove the noise. We use a ten sample wide moving average as the low pass filter and calculated the variance which is used in the activity detection stage. Variance is categorized by an increase in energy of the data. Since the variance is proportional to  $\sum(x^2) - (\bar{x}^2)$ , this was done efficiently

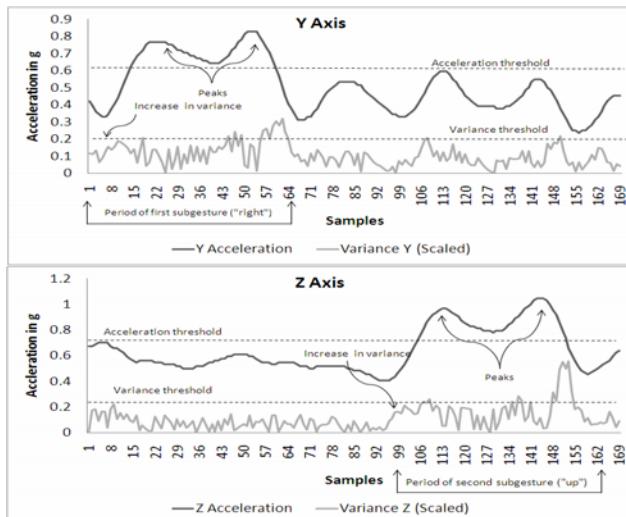
by also keeping a running sum of squares of the data. Ranges where the variance is greater than a set threshold are considered to be periods of activity.

Subgesture recognition is established using the following algorithm. A straight line motion in any axis will produce a two-peaked trace as shown in figure 3. After the motion is detected on the axis, it then counts the number of peaks in that direction. During the beginning of a subgesture, an increase in acceleration will cause the first peak, while at the stop of the subgesture, the deceleration ill cause the second peak. If the number of contained peaks is greater or less than the number of peaks required for a straight line before a new activity is detected, the previous subgesture is discarded.

Since the velocity of the arm is zero at the ends of the gesture, the integral of the acceleration across it must be zero as well. Completion detection is accomplished simply by tracking across an area of activity, and recording the number of peaks and their integral. A time limit is given to the user to begin the next gesture otherwise the current combination of subgestures is transferred to the computer to execute.

## 4 Experimental Results

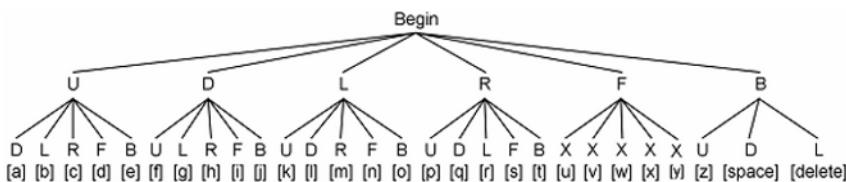
Figure 3 below shows the output of the system when an “up” + “right” gesture is executed. The subgestures are correctly extracted from the readings. The threshold value for the variance was found analytically by analyzing the data stream. A minimum peak size is assumed to reject noise and dithering.



**Fig. 3.** Results of a “right” + “up” gesture. The increase in variance sets of the activity detection stage. Note the number of peaks per subgesture.

Measurements of dynamic human gestures are inherently noisy as there is no clear distinction between when a subgesture stops and another subgesture begins. The system is incapable of detecting gestures like “forward” + “forward” for this same reason.

A decision tree for an application where gestures are assigned to the letters of the English alphabet is shown in figure 4. Only gestures of complexity two needs to be considered. For the output of figure 3, a character of ‘p’ was displayed by the system, corresponding to a “right” + “up” gesture.



**Fig. 4.** Decision tree showing outcomes of an application where gestures of complexity two is assigned a corresponding letter. D=Down, U=Up, L=Left, R=Right, F=Front, B=Back.

## 5 Conclusions and Future Direction

We have demonstrated that with a single 3 – axis accelerometer based wireless embedded system; we can distinguish various hand gestures of the user and by compounding them, more complex gestures can be realized. The hierarchy allows for future increase in the number of subgesture definitions, thereby increasing the number of gestures available to the user.

A single real life example showing the systems capabilities was demonstrated. The system has shown to prove reliable results. The design is simple and no calibration needs to be done by the user. We are currently investigating various wearable form factors of the system and increasing the array of subgestures that can be recognized.

In the short term, we plan to expand the definition of subgestures to include terms of magnitude and duration. Distance and velocity could also be used for precise control by integrating the acceleration values from the sensor. We also plan to increase the number of gestures recognized to utilize the full potential three dimensional space available.

In the long term, we envision our system to be used over a wide variety of applications, like annotating a data stream [3]. More intriguing is the concept of a generalized gesture recognition system, which switches from application to application simply by attaching it to a new object and configuring a new set of gestures and corresponding decision tree on the computer.

## References

1. Pavlovic, V.I., Sharma, R., Huang, T.S.: Visual interpretation of hand gestures for human-computer interaction: a review. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695 (1997)
2. Craven, M.P., Curtis, K.M., Hayes-Gill, B.R., Thursfield, C.D.: A Hybrid Neural Network/Rule-Based Technique for On-Line Gesture and Hand-Written Character Recognition. In: *Proceedings of the Fourth IEEE Int. Conf. on Electronics, Circuits and Systems*, vol. 2, pp. 850–853 (1997)
3. Chambers, G.S., Venkatesh, S., West, G.A.W., Bui, H.H.: Hierarchical recognition of intentional human gestures for sports video annotation. In: *Proceedings. 16th International Conference on Pattern Recognition*, vol. 2, pp. 1082–1085 (2002)
4. Cheok, A.D., Ganesh Kumar, K., Prince, S.: Micro-accelerometer based hardware interfaces for wearable computer mixed reality applications. In: *Proceedings. Sixth International Symposium on Wearable Computers*, pp. 223–230 (2002)
5. The I<sup>2</sup>C-BUS Specification Version 2.1 January 2000,  
[http://www.nxp.com/acrobat\\_download/literature/9398/39340011.pdf](http://www.nxp.com/acrobat_download/literature/9398/39340011.pdf)
6. Regenbrecht, H., Baratoff, G., Poupyrev, I., Billinghurst, M.: A Cable-less Interaction Device for AR and VR Environments. In: *Proceedings of ISMR*, pp. 151–152 (2001)
7. Perng, J.K., Fisher, B., Hollar, S., Pister, K.S.J.: Acceleration sensing glove (ASG). In: *The International Symposium on Wearable Computers. Digest of Papers*, pp. 178–180 (1999)
8. Reifinger, S., Wallhoff, F., Ablassmeier, M., Poitschke, T., Rigoll, G.: Static and Dynamic Hand-Gesture Recognition for Augmented Reality Applications. In: *Human-Computer Interaction. HCI Intelligent Multimodal Interaction Environments. LNCS*, vol. 4552, pp. 728–737. Springer, Heidelberg (2007)
9. Lam, A.H.F., Li, W.J.: MIDS: GUI and TUI in mid-air using MEP sensors. In: *The 2002 International Conference on Control and Automation. Final Program and Book of Abstracts*, pp. 86–87 (2002)
10. Van Laerhoven, K., Cakmakci, O.: What shall we teach our pants? In: *The Fourth International Symposium on Wearable Computers*, pp. 77–83 (2000)
11. Pylvänen, T.: Accelerometer Based Gesture Recognition Using Continuous HMMs. In: *Pattern Recognition and Image. LNCS*, vol. 3522, pp. 639–646. Springer, Heidelberg (2005)
12. Paradiso, J.A., Benbasat, A.Y.: An Inertial Measurement Framework for Gesture Recognition and Applications. In: *Gesture and Sign Language in Human-Computer Interaction. LNCS*, vol. 2298, pp. 77–90. Springer, Heidelberg (2002)
13. Deng, J.W., Tsui, H.T.: An HMM-based approach for gesture segmentation and recognition. In: *Proceedings. 15th International Conference on Pattern Recognition*, vol. 3, pp. 679–682 (2000)
14. Pengyu Hong, Turk, M., Huang, T.S.: Constructing finite state machines for fast gesture recognition. In: *Proceedings. 15th International Conference on Pattern Recognition*, vol. 3, pp. 691–694 (2000)
15. Bluetooth Technology, <http://www.bluetooth.com/Bluetooth/Technology/>
16. Wilson, A.D., Bobick, A.F.: Nonlinear PHMMs for the interpretation of parameterized gesture. In: *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 879–884 (1998)

17. LIS3LV02DQ Datasheet, MEMS Inertial Sensor, 3-Axis -  $\pm 2g/\pm 6g$  Digital Output Low Voltage Linear Accelerometer,  
<http://www.st.com/stonline/products/literature/ds/11115.pdf>
18. CY8C29466 PSoC® Mixed-Signal Array Datasheet, [http://download.cypress.com.edgesuite.net/design\\_resources/datasheets/contents/cy8c29466\\_8.pdf](http://download.cypress.com.edgesuite.net/design_resources/datasheets/contents/cy8c29466_8.pdf)
19. Parani-ESD100/110/200/210 User Guide and Datasheet,  
[http://www.sena.com/download/manual/manual\\_parani\\_esd-v1.1.4.pdf](http://www.sena.com/download/manual/manual_parani_esd-v1.1.4.pdf)

# **What Combinations of Contents is Driving Popularity in IPTV-based Social Networks?**

Rajen Bhatt

Samsung India Software R&D Center, Logix Infotech Park, D-5, Sector-59,  
Noida-201301, India  
Email: {rajen.bhatt@samsung.com, rajen.bhatt@gmail.com}

**Abstract.** IPTV-based Social Networks are gaining popularity with TV programs coming over IP connection and internet like applications available on home TV. One such application is rating TV programs over some predefined genres. In this paper, we suggest an approach for building a recommender system to be used by content distributors, publishers, and motion pictures producers-directors to decide on what combinations of contents may drive popularity or unpopularity. This may be used then for creating a proper mixture of media contents which can drive high popularity. This may also be used for the purpose of catering customized contents for group of users whose taste is similar and thus combinations of contents driving popularity for a certain group is also similar. We use a novel approach for this formulation utilizing fuzzy decision trees. Computational experiments performed over real-world program review database shows that the proposed approach is very efficient towards understanding of the content combinations.

## **1 Introduction**

Social networking websites are deriving popularity day-by-day. With television programs now coming over IP (or internet) connections, various social networking applications are becoming familiar sight at homes. These social networking applications offer various services on IPTV such as to identify presence of ones friends, family, and co-workers on IPTV network and what programs they are currently watching, watched or will watch in near future [1]. For example (taken from [1]), a client may determine that his friend John is watching a program entitled “Sweet Home Alabama”. The client may have the same program taste as John and therefore may decide to watch same program. If John is the favorite peer of the client, the program being watched by John may automatically play on client’s IPTV. Further client can query various other information such as peers’ recommendations, program ratings and reviews etc...Client can also make groups according to various tastes of programs such as Horror film group, Action film group and these groups can be constructed automatically based on the mining of user watch patterns. A method for sensing user presence for buddy list applications can be implemented on IPTV client in order to sense and display

presence information relating to a user watching television [2]. Rating service and recommendations compiles and processes ratings / reviews given by various viewers in the IPTV user group and provide recommendations to the viewers of the group based on the rating system [3]. On request from clients, it determine programs rated for viewing and rated not for viewing based on the content ratings. It can also initiate a recording of a recommended program with a television-based client device for a viewer. IPTV services are also available on computer with ip.tv [4,5]. It allows configuration of IPTV over a computer. It allows arranging conference calls, conferene recording, public chat, sharing status information, multi-conference, P2P conversations, transmitting camera captured video and movies, transmitting applications and application sharing, sharing white boards, creating polls, and sending files. Various other research and products related to IPTV are described in [6, 7, 8, 9].

The work presented in this research is towards a better method for representing user behavior so that understanding of contents, driving popularity can be understood. We develop human interpretable fuzzy rules over user ratings. These rules combine program genres with fuzzy connectives and generate decisions over Like and Dislike by a group of users over social network.

This paper is organized as follows. In Section II, we introduce notations and in parallel to that explain the proposed approach of generating the popularity driving fuzzy model. Computational experiments have been presented in Section III. Section IV concludes the paper with directions for further research.

## 2 The Proposed Content Popularity Modeling Approach

Let, a set of  $n$  programs  $\{M_1, M_2, \dots, M_n\}$  be described by a set of  $p$  genres  $\{g_1, g_2, \dots, g_p\}$ . Each genre  $g_j; j=1,2,\dots,p$  describe content of program like *Drama*, *Comedy* (or *Laughter*), *Epic*, *Suspense* etc...and restricted to lie in the range [0 5], i.e.,  $0 \leq g_j \leq 5, \forall j = 1, \dots, p$ . Let, each program  $M_i$  is reviewed by  $N_i$  IPTV viewers over  $p$  genres. Reviews collected over  $p$  genres for each program has been averaged out to generate the single rating vector for the program. If review vector by  $r^{\text{th}}$  reviewer for  $i^{\text{th}}$  program over  $p$  genres is given by  $\mathbf{R}^{ri}$  then average review vector for  $i^{\text{th}}$  program is given by

$$\mathbf{Re}^i = \frac{\sum_{r=1}^{N_i} \mathbf{R}^{ri}}{N_i} \quad (1)$$

In addition to described by  $p$  genres, each program  $M_i$  has been rated in one of the two categories  $\{\text{Like}, \text{Dislike}\}$  by each reviewer. For  $i^{\text{th}}$  program, category assigned by  $r^{\text{th}}$  reviewer is denoted as  $cat^{ri}$ . It is straight forward to mention that

$$cat^{ir} \in \{Like, Dislike\}, i = 1, \dots, n, r = 1, \dots, N_i.$$

Let  $N_{Li}$  and  $N_{Si}$  are number of users who have rated program  $M_i$  as *Like* and *Dislike*, respectively. Precisely,

$$\begin{aligned} N_{Li} &= \|cat^{ir}\|, cat^{ir} = \{Like\}, \\ N_{Si} &= \|cat^{ir}\|, cat^{ir} = \{Dislike\}. \end{aligned} \quad (2)$$

The overall rating for program  $M_i$  is fuzzy number with degrees of membership in *Like* and *Dislike* classes is given by the 2-element row vector

$$\mathbf{d}^i = \left[ \begin{array}{cc} N_{Li} & N_{Si} \\ \hline N_i & N_i \end{array} \right] = \left[ \begin{array}{cc} d^{Li} & d^{Si} \end{array} \right] \quad (3)$$

Here  $d^{Li}$  and  $d^{Si}$  are referred as the degrees with which  $i^{\text{th}}$  program has been categorized as *Like* or *Dislike*. Collection of  $d^{Li}$  and  $d^{Si}$  for  $i=1, \dots, n$  gives fuzzy class *Like* and *Dislike* referred here as  $L$  and  $S$ , respectively. This formulation of assigning fuzzy class for each program fulfills the FCM algorithm constraint of sum of degrees of membership of a pattern to  $q$  arbitrary fuzzy clusters should be one [10], *i.e.*,

$$\sum_{q=1}^k u_{iq} = 1, \forall i = 1, \dots, n. \quad (4)$$

With these notations, our program review data for all the programs contain  $p$ -dimensional review  $\mathbf{Re}^i; i=1, \dots, N_i$  and 2-dimensional overall rating memberships  $\mathbf{d}^i; i=1, \dots, N_i$ . Using this training data matrix, we generate fuzzy decision tree using fuzzy ID3 algorithm for the proposed recommender system. The entire recommender system has been represented by rules extracted from the fuzzy decision tree in a concise form.

We cluster the review data  $\mathbf{Re}^i; i=1, \dots, N_i$  into three fuzzy clusters using fuzzy  $c$ -means clustering algorithm. After clustering, each genre  $g_j$  has been represented by a term set  $Term(g_j) = \{F_{jk} | k = 1, \dots, c_j\}$ ;  $F_{jk}$  refers to  $k^{\text{th}}$  fuzzy set of  $g_j$ , and  $c_j$  is equal to number of fuzzy sets on  $g_j$ , which are taken as three here.  $\mu_{F_{jk}}(g_j^i)$  is membership degree of the  $i^{\text{th}}$  value of attribute  $g_j$  on the fuzzy set  $F_{jk}$ .

A typical fuzzy classification rule we generate using fuzzy ID3 algorithm can be written in the form:

“IF ( $g_1$  is  $F_{1m}$ ) AND ... AND ( $g_p$  is  $F_{pm}$ )  
THEN  $y'$  is  $L(\beta_{ml})$ ,  $y'$  is  $S(\beta_{ms})$ ”.

$y'$  is output of fuzzy classifier.  $\beta_{ml}$  ( $\beta_{ms}$ ) is referred as the certainty factor with which  $m^{\text{th}}$  rule predict program categorization class  $L$  ( $S$ ). In fuzzy classification problems, a *fuzzy evidence* is a conditional fuzzy set defined on the premise space, which represents the fuzzy values taken by one or more fuzzy attributes present in the premise part of the rule [11]. For the rule given above, fuzzy evidence  $F_v$  is a  $p$ -dimensional composite fuzzy set generated by  $F_{1m} \cap F_{2m} \cap \dots \cap F_{pm}$ , representing the condition that “IF ( $g_1$  is  $F_{1m}$ ) AND ... AND ( $g_p$  is  $F_{pm}$ )”. The *certainty factor* of a rule concerning classes  $L$  ( $S$ ) can be measured by fuzzy subsethood of  $F_v$  into  $L$  ( $S$ ), which is given by

$$\beta_{ml} = \frac{\sum_{i=1}^n \mu_{F_{1m}}(x_1^i) \cdot \mu_{F_{2m}}(x_2^i) \cdot \dots \cdot \mu_{F_{pm}}(x_p^i) d^{Li}}{\sum_{i=1}^n \mu_{F_{1m}}(x_1^i) \cdot \mu_{F_{2m}}(x_2^i) \cdot \dots \cdot \mu_{F_{pm}}(x_p^i)} \quad (5)$$

$\beta_{ml}$  and  $\beta_{ms}$  are degree of certainties with which  $m^{\text{th}}$  rule can classify arbitrary program in class *Like* or class *Dislike*. In certain cases, fuzzy rule may not use all the  $p$  genres  $\{g_1, g_2, \dots, g_p\}$ , but only a subset of it. Premise spaces of such rules are known as *partial premise space*. Rules constructed over partial premise space have better interpretability and lesser computational effort during inference.

#### A. Fuzzy Decision Trees and Fuzzy ID3

Fuzzy decision trees are composed of a set of internal nodes representing variables used in the solution of a classification problem, a set of branches representing fuzzy sets of corresponding node variables, and a set of leaf nodes representing the degree of certainty with which each class has been approximated.

Fuzzy ID3 is one of the algorithms available in the literature for the construction of fuzzy decision trees [11, 12, 13, 14]. Fuzzy ID3 is fuzzy version of ID3 algorithm initially proposed by Quinlan [15-24]. Fuzzy ID3 utilizes fuzzy classification entropy of a possibilistic distribution for decision tree generation. For each fuzzy set  $\{F_{jk} | k=1, \dots, c_j\}$  of the genre  $g_j$ , certainty factor concerning the class  $L$  is defined as

$$\beta_{jk}^L = \frac{\sum_{i=1}^n \min(\mu_{F_{jk}}(x_j^i), d^{Li})}{\sum_{i=1}^n \mu_{F_{jk}}(x_j^i)}; 0 \leq \beta_{jk}^L \leq 1. \quad (6)$$

$\beta_{jk}^S$  can be calculated using the same formula by replacing  $d^{Li}$  by  $d^{Si}$ .

**Definition 1:** The fuzzy classification entropy of  $F_{jk}$  is defined as

$$Entr_{jk} = \sum_{class \in \{L, S\}} \beta_{jk}^{class} \times \log_2 (\beta_{jk}^{class}) \quad (7)$$

With  $0 \leq \beta_{jk}^{class} \leq 1$ , the function  $Entr_{jk}$  attains its minimum at a 2-dimensional vector of components  $\beta_{jk}^{class} (class \in \{L, S\})$  with each component either 0 or 1 (for proof see Ref. [14]), with the assignment

$$0 \times \log_2 0 = \lim_{\beta_{jk}^{class} \rightarrow 0} \beta_{jk}^{class} \times \log_2 \beta_{jk}^{class}. \quad (8)$$

**Definition 2:** The averaged fuzzy classification entropy of genre  $g_j$  is defined as

$$E_j = \sum_{k=1}^{c_j} w_{jk} \times Entr_{jk}. \quad (9)$$

$w_{jk}$  denotes the weight of the  $k^{\text{th}}$  fuzzy set of  $j^{\text{th}}$  attribute and is defined as

$$w_{jk} = \frac{\sum_{i=1}^n \mu_{F_{jk}}(x_j^i)}{\sum_{k=1}^{c_j} \left( \sum_{i=1}^n \mu_{F_{jk}}(x_j^i) \right)}. \quad (10)$$

Given the fuzzy partition of feature space, leaf selection threshold  $\beta_m$ , and expanded attribute selection criterion given in Eq. (9), the general procedure for generating fuzzy decision tree using fuzzy ID3 is outlined as follows:

**WHILE** there exist candidate nodes

**DO** select one attribute using the search strategy. Generate its child-nodes according to an expanded attribute obtained by the given heuristic. Check child-nodes for the leaf selection threshold. Child-nodes meeting the threshold have to be terminated as leaf nodes. The remaining child-nodes are regarded as new candidate nodes.

### 3 Computational Experiments

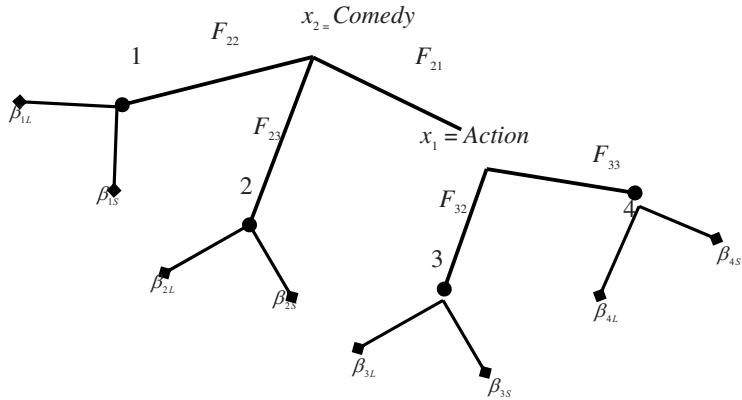
We demonstrate the proposed approach using the example dataset given in Table I. We have considered 10 televisions programs over four common genres *Action*, *Comedy*, *Drama*, and *Musical*. Each program has been reviewed by at least 5 reviewers. Table I given below gives the consolidated average ratings.  $d^{Li}$  and  $d^{Si}$  have been calculated as described in Section II.

Figure 1 shows fuzzy decision tree using fuzzy ID3 algorithm for the ratings given in Table I. In Fig. 1, root node is represented by  $g_2$  (*i.e.*, *Comedy*). There are total four leaf nodes shown by bold dots and marked as 1,...,4. Review entered for new programs are classified by starting from the root node and then reaching to one or more than one leaf nodes by following the path of degree of membership greater than zero.

**Table 1.** consolidated dataset for the example movie recommender system problem

Index	Considered genres				Rating	
	Action $g_1$	Comedy $g_2$	Drama $g_3$	Musical $g_4$	$d^{Li}$	$d^{Si}$
$M_1$	7	5	5	4	0.90	0.10
$M_2$	1	7	6	8	0.98	0.02
$M_3$	0	7	3	3	0.3	0.70
$M_4$	0	4	6	8	0.35	0.65
$M_5$	2	5	5	9	0.85	0.15
$M_6$	0	4	8	7	0.95	0.05
$M_7$	7	2	6	0	0.4	0.60
$M_8$	1	4	7	9	0.25	0.75
$M_9$	3	1	3	5	0.15	0.85
$M_{10}$	2	3	2	6	0.10	0.90

Four fuzzy classification rules shown below can be extracted from the fuzzy decision tree shown in Fig. 3.



**Fig. 3** Fuzzy Decision Tree for Example Recommender System Problem

If (Comedy is  $F_{22}$ , i.e., High) Then  $y' = Like (\beta_{1L})$  and  $y' = Dislike (\beta_{1S})$

If (Comedy is  $F_{23}$ , i.e., Low) Then  $y' = Like (\beta_{2L})$  and  $y' = Dislike (\beta_{2S})$

If (Comedy is  $F_{21}$ , i.e., Medium) AND (Action is  $F_{32}$ , i.e., Medium) Then  $y' = Like (\beta_{3L})$  and  $y' = Dislike (\beta_{3S})$

If (Comedy is  $F_{21}$ , i.e., Medium) AND (Action is  $F_{33}$ , i.e., High) Then  $y' = Like (\beta_{4L})$  and  $y' = Dislike (\beta_{4S})$

Certainty factors associated with four leaf nodes for both the classes are

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1L} & \beta_{1S} \\ \beta_{2L} & \beta_{2S} \\ \beta_{3L} & \beta_{3S} \\ \beta_{4L} & \beta_{4S} \end{bmatrix} = \begin{bmatrix} 0.04 & 0.95 \\ 0.97 & 0.02 \\ 0.91 & 0.08 \\ 0.02 & 0.97 \end{bmatrix}.$$

The interpretation of above rule base along with certainty factors clearly indicates that programs having low comedy content or medium comedy content with medium action content are more likely to get the popularity in certain group

of users. On the other front, High comedy content or medium comedy content with high action content may damage the ratings and thus popularity of the program. The above rule base also provides interpretation that drama and musical contents are actually not affecting the popularity of programs in a certain social network. So broadcasters may keep focus on releasing the content which is having proper popularity driving combinations of comedy and action and just avoid the drama and musical programs.

Further, if program directors / producers wants to compile and produce completely a new program for catering to certain social network they can refer to fuzzy model one like the given above in Fig. 1 along with the extracted fuzzy rules. This gives an idea of controlling the contents in the program so that the new program is more likely to get the popularity within certain social network. For a newly developed content one can actually predict the degrees of membership into *Like* and *Dislike* classes using the fuzzy inference mechanism.

Additionally advertisers can use this fuzzy model to decide on the types of advertisements to be delivered during the program breaks. For example, advertisers can deliver a funny mobile device advertisement on a certain social network which likes medium comedy content or action pack high power bike advertisement on a certain social network which likes medium action contents as well.

To demonstrate the further effectiveness of the proposed framework, we have conducted the exhausted computational experiments over 150 programs reviewed by 580 users. Each program has reviewed more than one program. The fuzzy decision tree and fuzzy rule base created on this 150 patterns program review data gives excellent results for content combinations and prediction of overall program recommendations. We are planning to build the proposed concept within IPTV network as a recommender system for the viewers, broadcasters, and producers.

## 4 Conclusion and Further Work

In this paper, we have presented an approach to construct a fuzzy model describing the contents driving popularity and dispopularity within certain social network. The approach described is very intuitive and computational complexity involved is only of the order of number of programs broadcasted or over which this model is required to be built. This makes the approach easily configurable and portable to existing x86 based digital television hardware. The model can be used by the broadcasters for catering the content, advertisers to identify the types of advertisements, viewers to understand the behavior of buddies, and by producers / directors to create the right combinations which can drive popularity over society. We are planning to build the proposed approach within IPTV network doing the rigorous experiments over user of various age, gender, profession, and ethnicity. This may explore new directions for us in understanding the content tastes across the society and develop / deliver the customized contents for Video-on-Demand framework.

## References

1. US Patent no-US 2007/0078971, Methods, systems, and computer program products for providing activity data, April 5, (2007)
2. US Patent no-US 2007/0288627, "Method for sensing user presence for buddy list applications", December 13, (2007)
3. US Patent no-US 2007/0204287, Microsoft Corporation, "Content ratings and recommendations", August 30, (2007)
4. <http://www.ip.tv/download>
5. Digital convergence of ideas and people, ip.tv users' guide, version 3.15.42.331, <http://www.ip.tv/download/userguide.pdf>.
6. US Patent no-US 2007/0277205, SBC Knowledge Ventures L.P., "System and method for distributing video data", November 29, (2007)
7. US Patent no-US 2008/0092199, SBC Knowledge Ventures L.P., "System and method for distributing dynamic event data in an internet protocol television system", April 17, (2008)
8. US Patent no-US 2008/0040742, SBC Knowledge Ventures L.P., "Method and system for inserting advertisement data into an internet protocol television network", February 14, (2008)
9. US Patent no-US 2008/0059998, SBC Knowledge Ventures L.P., "System and method of communicating emergency alerts", March 6, (2008)
10. N.R. Pal and J.C. Bezdek, "On cluster validity for fuzzy  $c$ -means model", IEEE Transactions On Fuzzy Systems 3, 370–379 (1995)
11. Y. Yuan and M.J.Shaw, "Induction of fuzzy decision trees", Fuzzy Sets and Systems, 69, 125–139, (1995)
12. Rajen B. Bhatt, Fuzzy-rough approach to pattern classification: hybrid algorithms and optimization, PhD Thesis, Electrical Engineering Department, IIT Delhi, India. Call number: 621-52 BHA-F, Accession no: TH-3259, [http://indest.iitd.ac.in/scripts/wwwi32.exe/\[in=arphd\]/](http://indest.iitd.ac.in/scripts/wwwi32.exe/[in=arphd]/).
13. Bhatt R.B.: Fuzzy Deicision Trees: Approaches, Advances, and Applications, In: Proceedings of National Conference on Security and Soft Computing (NSSC), pp. 6-10, March 29-31, Surat, India, (2007)
14. Wang X.-Z., Yeung D.S., Tsang E.C.C.: A comparative study on heuristic algorithms for generating fuzzy decision trees, IEEE Transactions on SMC – B 21, 215–226 (2001)
15. Quinlan J.R, Induction of decision trees, Machine Learning 1, 81–106 (1986)
16. Umano M. *et.al.*, "Fuzzy decision tree by fuzzy ID3 algorithm and its application to diagnosis systems", in: Proceedings of IEEE International Conference on Fuzzy Systems, pp.2113-2118, June 26-29, (1994)
17. Chiang I.-J., jen Hsu J.Y.-.: Fuzzy classification trees for data analysis, Fuzzy Sets and Systems 130, 87–99 (2002)
18. Jeng B., Jeng. Y.-M., Liang T.-P.: FILM:A fuzzy inductive learning method for automated knowledge acquisition, Decision Support Systems 21, 61–73 (1997)
19. Ichihashi H., Shirai T., Nagasaka K., Miyoshi, T. : Neuro-fuzzy ID3, Fuzzy Sets and Systems . 81, 157–167 (1996)
20. Sison L.G. Chong, E.K.P.: Fuzzy modeling by induction and pruning of decision trees, in: Proceedings of IEEE International Symposium on Intelligent Control, pp. 166-171, Columbus, OH, Aug. (1994)
21. Tani T., Sakoda M.: Fuzzy modeling by ID3 algorithm and its application to prediction of outlet temperature, In: Proceedings of IEEE International Conference on Fuzzy Systems, pp. 923–930, San Diego, CA, (1992)

22. Weber R.: Fuzzy ID3: a class of methods for automatic knowledge acquisition, In: Proceedings of International Conference on Fuzzy Logic and Neural Networks, pp. 265–268., Iizuka, Japan (1992).
23. Mitra S., Knowar K.M, Pal S.K.: Fuzzy Decision Tree, Linguistic Rules and Fuzzy Knowledge-Based Network: Generation and Evaluation, IEEE Transactions on SMC-C: Applications and Reviews. 32, 328–339 (2002)
24. Janikow C.Z.,: Fuzzy Decision Trees: Issues and Methods. IEEE Transactions on SMC-B: Cybernetics 28, 1–14 (1998)

# Adaptive Acceleration of MAP with Entropy Prior and Flux Conservation for Image Deblurring

Manoj Kumar Singh<sup>1</sup>, Yong-Hoon Kim<sup>2\*</sup>, U. S. Tiwary<sup>3</sup>, Rajkishore Prasad<sup>4</sup>  
and Tanveer Siddique<sup>3</sup>

<sup>1</sup> Dept. of Computer Science and Engineering, Galgotiya College of Engineering and Technology, India

<sup>2</sup>Sensor System Laboratory, Department of Mechatronics, Gwangju Institute of Science and Technology (GIST), Republic of Korea

<sup>3</sup>Indian Institute of Information Technology Allahabad(IIITA), India

<sup>4</sup>University of Electro-Communication, Tokyo, Japan

<sup>1</sup>E-mail: mks\_kjist@yahoo.co.in, <sup>2</sup>\*E-mail: yhkim@gist.ac.kr

**Abstract:** In this paper we present an adaptive method for accelerating conventional *Maximum a Posteriori* (MAP) with Entropy prior (MAPE) method for restoration of an original image from its blurred and noisy version. MAPE method is nonlinear and its convergence is very slow. We present a new method to accelerate the MAPE algorithm by using an exponent on the correction ratio. In this method the exponent is computed adaptively in each iteration, using first-order derivatives of deblurred images in previous two iterations. The exponent obtained so in the proposed accelerated MAPE algorithm emphasizes speed at the beginning stages and stability at later stages. In the accelerated MAPE algorithm the non-negativity is automatically ensured and also conservation of flux without additional computation. The proposed accelerated MAPE algorithm gives better results in terms of RMSE, SNR, moreover, it takes 46% lesser iterations than conventional MAPE.

## 1 Introduction

Image deblurring is a longstanding linear inverse problem and is encountered in many application areas such as remote sensing, medical imaging, seismology, and astronomy [1, 2, 3]. Generally, many linear inverse problems are ill-conditioned, since either the inverse of linear operators does not exist or is nearly singular yielding highly noise sensitive solutions. Most of the methods given to solve ill-conditioned problems are classified into following two categories: a) Methods based on regularization [2, 3] and b) Methods based on Bayesian theory [1, 2, 4–8].

The centric idea in regularization and Bayesian approaches is the use of *a priori* information expressed by regularization/ prior term. A good prior term gives a higher score to most likely images. However, modeling a prior for real-word images is subjective matter and is not trivial. Many directions for prior modeling have been proposed such as derivative energy in the Wiener filter [2, 3] compound Gauss Markov random field [2, 13], Markov random fields with non quadratic potentials [2, 11, 13], Entropy [1, 8, 10], and heavy tailed densities of images in wavelet domain [12]. However, in the absence of any prior/preliminary information about the original image, entropy is considered as the best choice to define prior term [4].

MAPE algorithm developed under Bayesian framework is nonlinear and solved iteratively [8, 10]. However, it has the drawbacks of slow convergence and being computationally expensive. Many techniques for accelerating the iterative method have been proposed and these can also be used for accelerating the MAPE algorithm [1, 9]. All these methods use correction terms - may be negative at times – which are computed in every iteration, multiplied with acceleration parameter, and added to the results obtained in previous iteration. Because the correction term may be negative at times, the non-negativity of pixel intensity in restored image is not guaranteed. In these acceleration methods positivity is enforced manually at the end of iterations. The main drawback of these acceleration methods is the selection of optimal acceleration parameter. Large acceleration parameter speeds up the algorithm, but it may introduce error. If error is amplified during iteration, it can lead to instability. Thus these methods require a correction procedure in order to ensure the stability. This correction procedure reduces the gain obtained by acceleration step and also needs extra computation.

In this paper we propose a new adaptive acceleration method for MAPE algorithm in order to cope with the problems of earlier acceleration methods. The proposed acceleration method requires minimum information about the iterative process. We use an exponent on multiplicative correction as an acceleration parameter which is computed adaptively in each iteration using the first order derivative of deblurred image from previous two iterations. The positivity of pixel intensity in the proposed acceleration method is automatically ensured since multiplicative correction term is always positive. Maintaining the total flux is important for applications where the blurring does not change the total number of photons or electrons detected. In this method we also achieve flux conservation without extra computational overhead. Section 2 discusses the conventional MAPE and accelerated MAPE algorithm. Section 3 describes the adaptive selection of an acceleration parameter. In Section IV experimental setup and results are presented. Section V gives the conclusion, which is followed by references.

## 2 Accelerated MAPE Algorithm with Flux Conservation

Let an original image, size  $M \times N$  blurred by shift-invariant point spread function (PSF) and corrupted by Poisson noise. This can be written in matrix form as [14]:

$$y = Hx + n, \quad (1)$$

where  $H$  is  $MN \times MN$  block Toeplitz matrix representing a linear shift-invariant PSF;  $x$ ,  $y$ , and  $n$  are vectors of size  $MN \times 1$  containing the original image, observed image, and sample of noise, respectively, arranged in column lexicographic ordering. The aim in image deblurring is to find an estimate of an original image  $x$  for a given blurred image  $y$  blurring operator  $H$  and distribution of noise  $n$ .

We derive the MAPE algorithm, in Bayesian framework, with Poisson type noise  $n$ . The basic idea of Bayesian framework is to incorporate the prior information, about the solution. A prior information is included using *a priori* distribution. In MAPE algorithm, *a priori* distribution,  $p(x)$ , is defined using entropy as:

$$p(x) = \exp(-E(x)), \quad (2)$$

where  $E(x)$  is the entropy of the original image  $x$ . We use the following entropy function:

$$E(x) = -\sum_i x_i \log x_i \quad (3)$$

When  $n$  is zero in Eq.(1), we consider only blurring, the expected value at the  $i^{th}$  pixel in the blurred image is  $\sum_j h_{ij} x_j$ . Where  $h_{ij}$  is  $(i, j)^{th}$  element of  $H$  and  $x_j$  is the  $j^{th}$  element of  $x$ . Because of Poisson noise, the actual  $i^{th}$  pixel value  $y_i$  in  $y$  is one realization of Poisson distribution with mean  $\sum_j h_{ij} x_j$ . Thus we have following relation:

$$p(y_i/x) = \left( \sum_j h_{ij} x_j \right)^{y_i} \exp\left(-\sum_j h_{ij} x_j\right) / y_i!. \quad (4)$$

Each pixel in blurred and noisy image,  $y$ , is realized by an independent Poisson process. Thus the likelihood of getting noisy and blurred image  $y$  is given by:

$$p(y/x) = \prod_i \left[ \left( \sum_j h_{ij} x_j \right)^{y_i} \exp\left(-\sum_j h_{ij} x_j\right) / y_i! \right]. \quad (5)$$

MAPE algorithm method with flux conservation for image deblurring, seeks an approximate solution of Eq. (1) that maximizes the a posteriori probability  $p(x/y)$  or  $\log p(x/y)$ , subject to the constraint of flux conservation,  $\sum_j x_j = N$ , where  $N$  is the sum of pixel values in observed image. We consider the maximization of following function:

$$L(x, \mu) = \log p(x/y) - \mu \left( \sum_j x_j - N \right). \quad (6)$$

Where  $\mu$  is the Lagrange multiplier for flux conservation. Now from Bye's theorem substitution of  $p(x/y)$  in terms of  $p(y/x)$  in Eq.(6), and then using  $p(x)$ ,  $p(y/x)$  from Eq.(2), (5) we get:

$$L(x, \mu) = \sum_i \left[ -\sum_j h_{ij} x_j + y_i \log \left( \sum_j h_{ij} x_j \right) \right] - \sum_j x_j \log x_j - \mu \left( \sum_j x_j - N \right). \quad (7)$$

For maximization of  $L$ ,  $\partial L(x, \mu) / \partial x_k = 0$ , we get the following relation:

$$1 + \mu = \sum_i \left[ h_{ik} \left\{ \left( y_i / \sum_j h_{ij} x_j \right) - 1 \right\} \right] - \log(x_k). \quad (8)$$

Eq.(8) is nonlinear in  $x_k$ , and is solved iteratively. By adding a positive constant  $C$  and raising exponent  $q$  both sides of Eq. (8), and then multiply both sides by  $x_k$ , we arrive at the following iterative procedure:

$$x_k^{l+1} = A x_k^l \left[ \sum_i \left( h_{ik} y_i / \sum_j h_{ij} x_j^l \right) - 1 + \log x_k^l + C \right]^q, \quad (9)$$

where  $A = [1 + \mu + C]^q$ . For ensuring the non-negativity of  $x_k^l$ , which allow the computation of  $\log x_k^l$  in the next iteration, a suitable constant  $C$  is selected. The constant  $A$  is recalculated at the end of each iteration using constraint  $\sum_j x_j^l = N$ . Accordingly, we get following:

$$A(l) = N \left[ \sum_k x_k^l \left\{ \sum_i \left( h_{ik} y_i / \sum_j h_{ij} x_j^l \right) - 1 - \log x_k^l + C \right\} \right]^{-q} \quad (10)$$

.

It has been found that the iteration given in Eq. (9) converges for  $1 \leq q \leq 3$ . Under such limit, the larger values of  $q$  give faster convergence but the risk of instability increases and the smaller values of  $q$  lead to slow convergence with improved stability. Thus, Eq. (9) with adaptive selection of an exponent  $q$  leads to the adaptively accelerated MAPE algorithm. For  $q = 1$ , Eq. (9) results into conventional MAPE algorithms.

### 3. Adaptive Selection of Exponent $q$

The choice of  $q$  in Eq. (9) mainly depends on the noise,  $n$ , and its amplification during iterations. If noise is high, a smaller value of  $q$  is selected and vice-versa. Thus the convergence speed of the proposed method depends on the choice of the parameter  $q$ . Drawback of this accelerated form of MAPE algorithm is that the selection of exponent  $q$  has to be done manually by trial and error. We overcome this serious limitation by proposing a method in which  $q$  is computed adaptively as iterations proceed. Proposed expression for  $q$  is as follows:

$$q(l+1) = \exp \left( \frac{\|\nabla x^l\|}{\|\nabla x^{l-1}\|} \right) - \frac{\|\nabla x^2\|}{\|\nabla x^1\|} \quad (11)$$

where  $\nabla x^l$  stands for the first-order derivative of  $x^l$  and  $\|\cdot\|$  denotes the L2 norm. Main idea in using first-order derivative is to utilize the sharpness of image. Because of the blurring, the image becomes smooth, sharpness decreases, and edges are lost or become weak. Deblurring makes image non-smooth, and

increases the sharpness. Hence the sharpness of deblurred image,  $\nabla x^l$ , increases as iterations proceed. For different levels of blur and different classes of images, it has been found by experiments that L2 norm of gradient ratio  $\|\nabla x^l\|/\|\nabla x^{l-1}\|$  converges to 1 as the number of iterations increase. Accelerated MAPE algorithm emphasizes speed at the beginning stages of iterations by forcing  $q$  around three. When the exponential term in Eq. (11) is greater than three, the second term,  $\|\nabla x^2\|/\|\nabla x^1\|$ , limits the value of  $q$  within three to prevent divergence. As iterations increase the second term forces  $q$  towards the value of one which leads to stability of iteration. By using the proposed exponent,  $q$ , the method emphasizes speed at the beginning stages and stability at later stages of iteration. Thus selecting  $q$  given by Eq. (11) for iterative solution Eq. (9) gives accelerated MAPE algorithm with adaptive selection of acceleration parameter. The non-negativity of pixel intensity is automatically ensured, since correction ratio Eq. (9) is always positive. In order to initialize the proposed method, first two iterations are computed using some fixed value of  $q$  ( $1 \leq q \leq 3$ ). In order to avoid the instability at the starting of the iteration,  $q = 1$  is preferable choice.

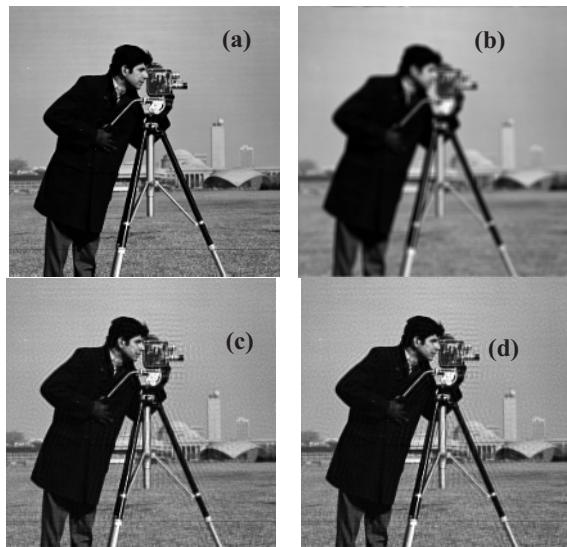
#### 4. Experiment and Results

For experiment, we choose the gray scale test image “Cameraman” (8-bit, 256  $\times$  256), uniform  $5 \times 5$  Box-car PSF, and Poisson noise. The blurred signal to noise ratio ( $BSNR$ ) is defined in decibel as below and was set to 40 dB [14].

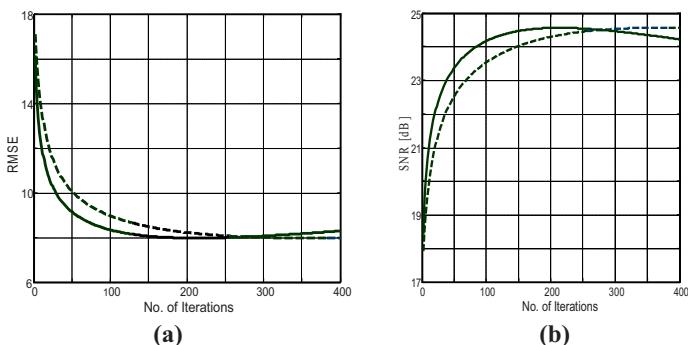
$$BSNR = 10 \log_{10} \left[ \sum \{ Hx - (1/MN) \sum Hx \}^2 / \sum (y - Hx)^2 \right] \quad (12)$$

The RMSE and SNR criteria are used for performance comparison of conventional and adaptive accelerated MAPE algorithm are defined as:

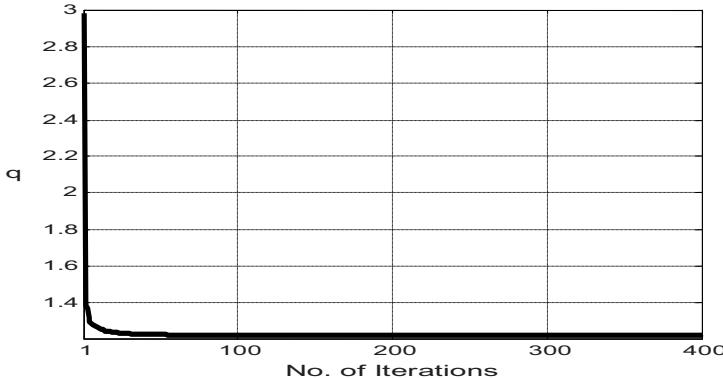
$$RMSE = \sqrt{(1/MN) \sum (y - x^k)^2}, \quad SNR = 10 \log_{10} \left( \sum |y|^2 / \sum |y - x^k|^2 \right) \quad (13)$$



**Fig. 1.** “Cameraman” a) Original b) Noisy and Blurred c) Restored by MEM corresponding maximum SNR in 367 iterations  
d) Retored Image by Accelerated MEM corresponding maximum SNR in 200 iterations



**Fig. 2.** a) RMSE of the MAPE (dotted line), RMSE of the Accelerated MAPE (solid line) b) SNR of the MAPE (dotted line), SNR of the Accelerated (solid line)



**Fig. 3.** Iterations Vs.  $q$

Figure 1 (c), (d) show the results of the MAPE and accelerated MAPE algorithm corresponding to maximum SNR. Figures 2 (a), (b), show the variation of SNR, RMSE versus iterations for MAPE and accelerated MAPE algorithm. It is observed that the accelerated MAPE algorithm has faster increase in SNR and faster decrease in RMSE in comparison of conventional MAPE algorithm. Figure 3 shows the variation of exponent  $q$  versus iteration number.

## 5. Conclusion

We have given a new method to accelerate the conventional MAPE algorithm. This method adaptively computes exponent of correction term in each iteration using the first-order derivative of the restored image in previous two iteration. From experiment, it is found that accelerated MAPE algorithm gives better results in terms of RMSE, high SNR, in 46% lesser iterations than the conventional MAPE algorithm. While computations required per iteration in MAPE as well as accelerated MAPE algorithms are almost same. This adaptive acceleration method has simple form and can be very easily implemented. Accelerated MAPE algorithm automatically preserves the non-negativity and flux, without additional computations.

## Acknowledgement

This work was supported by the Dual Use Center through the contract at Gwangju Institute of Science and Technology and by the BK21 program in Republic of Korea.

## References

1. Jansson P. A.: Deconvolution of Images and Spectra. New York: Academic Press; (1997)
2. Katsaggelos A.K.: Digital Image Restoration. New York: Springer-Verlag (1989)
3. A.K. Jain. Fundamental of Digital Image Processing. Engelwood Cliffs, NJ: Prentice-Hall (1989)
4. Stark J.L., Murtagh F., Querre P., Bonnarel F.: Entropy and astronomical data analysis. Perspectives from multiresolution. *A & A* 368, 730–746 (2001)
5. W. H. Richardson. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Amer.* 62(1), 55–59 (1972)
6. Lucy L. B. : An iterative techniques for the rectification of observed distributions. *Astronom. J.* 79(6), 745–754 (1974)
7. Meinel E. S. : Origins of linear and nonlinear recursive restoration algorithms. *J. Opt. Soc. Amer. A* 3(6), 787–799 (1986)
8. Meinel E. S.: Maximum entropy imaage restoration. *J. Opt. Soc. Amer. A* 5(1), 25–29. (1988)
9. Biggs D.S.C., Andrews M.: Acceleration of Iterative image restoration algorithms. *Applied Optics* 36(8),1766–1775 (1997)
10. Nunez J., Llacer J. : A fast Bayesian reconstruction algorithm for emission tomography with entropy prior converging to feasible images. *IEEE Trans. on Med. Imaging* 1990; 9( 2), 159–171 (1990)
11. Nikolova M. :Local strong homoginity of a regularized estimator. *SIAM J. Appl. Math.* 61, 3437–3450 (2000).
12. Dias J. : Bayesian Wavelet based image deconvolution: A GEM algorithm exploiting a class of heavy-tailed priors. *IEEE Trans Image Process* 15( 4), 937–951 (2006).
13. Zhou Z., Leahy M., Qi J. : Approximate maximum likelihood hyperparameter estimation for Gibbs priors. *IEEE Trans Image Process* 6(6), 844–861 (1997)
14. Katkovnic V., Egiazarian K., Astola J.: Local approximation techniques in signal and image processing. Bellingham, Washington: SPIE press (2006)

# Language Processing

# An Intelligent Automatic Text Summarizer

M. Shoaib Jameel, Anubhav, Nilesh Singh, Nitin Kumar Singh, Chingtham Tejbanta Singh and M.K. Ghose

Department of Computer Science and Engineering

Sikkim Manipal Institute of Technology,

Majitar, Rangpo, East Sikkim 737132. INDIA

{shoaib.jameel, anubhavsmi, nilesh.singh69, nitinchantu, chingtham,  
headcse.smit}@gmail.com

**Abstract.** This paper describes an intelligent text summarizer that summarizes a given piece of text into three different summaries based on three different algorithms. This summarizer uses statistical methods to summarize a text like considering the frequency of words, rare words etc. It then gives a meaningful title to the main text and finally selects the best summary out of a list of given summaries. This summarizer allots the writer a competence level (in written English) after analyzing the text like number of rare words used. The title generator of the summarizer gives a short title to the main text. Results obtained through experiments showed that it is indeed possible to determine the competence level of the writer from the text and proximity of the sentences play a vital role in selecting the best summary.

## 1 Introduction

A summary is a text that is produced from one or more texts, that contain a significant portion of information in the original text, and that is no longer than half the original text [1]. Summarization is used extensively in generating search engine query results and in generating the automated abstracts of research papers. Summarization plays an important role in categorizing the ever-growing extensive collection of web pages that is present in the web today [2] [3]. Title extraction [4] [5] is an important and the most challenging task for any summarizer. The most difficult part for any summarizer is choosing which sentences are important and need to be selected for final summarization. This summarizer has been developed keeping in mind that there is no loss of information during summarization phase.

## 2 Related Work

This work is a part of the ongoing research called The Thinking Algorithm [6], where the indexer does text summarization. According to [7] there are two ways to view text summarization either as text extraction or as text abstraction. This summarizer draws much inspiration from H.P. Luhn's work of text summarization [12]. What differentiates this work with the other works is that most of the works cited in this paper generate one summary, which is claimed to be the best one. This summarizer generates three different summaries. Moreover, the algorithms and heuristics applied in the summarizer that select the final summary take into account the proximity of the sentences in the main text compared to that generated in the summarized text. Moreover, none of the other works try to understand the competence level of the writer. Other works in text summarization both statistical and machine learning include [13] [14] [15] [16] [17]. In [8] a method called lexical aggregation is explained. Lexical aggregation is related to the so-called concept fusion described in [7]. There are three steps to perform text summarization. First, understand the topic of a text [7]; second the interpretation of the text and finally the generation of the text. The generation of text is carried out in two different ways, namely: Extraction and Abstraction. The abstraction generation must make use of a natural language generator as for example [9]. Method to identify topics is to perform rhetorical parsing and hence build RST tree where the nuclei is identified as the topic [10]. Multitext text summarization has been investigated in [11].

## 3 Implementation Details

Five persons were given the task of drafting an essay (of about 500 words) from a topic chosen unanimously. The writers were not equally competent in written English. Following were addressed during implementation:

- Why is it so that one person uses more rare words in the text as compared to the other person?
- Why is it so that one person can enumerate several ideas in just one long sentence while the other uses several sentences in order to convey the same idea?
- How can the competency of the writer be judged from the essay?
- Why has the writer chosen one particular word in a sentence though other words of the same meaning do exist?
- How can the title be given to each of the text, using an automatic text summarizer?
- How important are the Proper Nouns, in the text?
- How to determine which sentence is important in which case and which one needs to be selected by the automatic text summarizer?

## 4 Observations and Algorithm Design

Observation 1: The writer with excellent command in English wrote long sentences with selection of an appreciable amount of high-frequency words.

Observation 2: Two writers wrote average English essay, with less use of high-frequency words and a mixture of long and short sentences.

Observation 3: The remaining two writers were not so competent in English writing. They wrote essays with an appreciable amount of grammatical mistakes. Sentences written were very short.

It became clear that high-frequency words play an important role in any writing because they convey some message from the writer. A writer uses high-frequency words to express an important message. For example, one of the writers wrote: "*Students must never infringe the rubric of the school.*" Automatically, this becomes an important line because it conveys an important aspect, which should always be followed by the students. "*infringe*" and "*rubric*" were rare words<sup>1</sup> in the whole essay.

No loss of information from the main text was also considered an important part. It would be better if all the sentences of similar category were grouped together. Using a selection algorithm the best group could be chosen based on some rules.

Each word should be allotted some weight based on their occurrence in the main text. Weights can be equaled to the words' frequency i.e. how many times a word occurs in the text. After several evaluations and experiments, following three algorithms were chosen.

1. Extracting and grouping those sentences that have rare words in them.
2. Extracting those sentences whose sum of frequencies of words in that sentence is below the average value calculated from the entire sum of frequencies of words in the whole document.
3. Extracting those sentences whose sum of frequencies of words in that sentence is above the average value calculated from the entire sum of frequencies of words in the whole document.
4. Selection algorithm, the algorithm that selects one best summary out of a list of three summaries.

### 4.1 Explanation and Proof of Concepts

The sentences with rare words were categorized into one group. However, the problem with this algorithm was observed when it summarized that essay of the writer with excellent command in written English. The algorithm selected more than 90% of the lines of the essay. This violated from the definition of a summary that it should not be longer than half the original text.

Algorithms (2) and (3) stated above work as follows:

---

<sup>1</sup> Rare Words are those words that occur only once in the entire document.

Let there exist a sentence S in the main text. Let there be N number of words in the sentence with frequencies  $f_1, f_2, f_3 \dots f_N$ . This frequency is the total count of the occurrence of the words in the whole document. The sum of the frequencies ( $S_F$ ) is then calculated for that sentence i.e.  $f_1 + f_2 + f_3 + \dots + f_N$ . The average frequency ( $Avg_F$ ) is calculated from the formula listed in Fig. 1. If  $S_F < Avg_F$  i.e. algorithm (2) and if  $S_F > Avg_F$  i.e. algorithm (3).

Following algorithm was devised:

Step 1: Calculate frequencies of the words present in the main text.

Step 2: Find the sum of the all the frequencies and hence calculate the average of the frequencies.

$$\text{Average Frequency} = \frac{\text{Sum of frequencies of individual words}}{\text{Total number of words}}$$

Fig. 1. Calculating the average frequency ( $Avg_F$ ) of the whole document.

Step 3: Store the rare words in a separate file. Rare words are those words that are defined by a frequency of one.

Step 4: Find the sum of the frequencies of each sentence and compare the sum with the value obtained from Step 2, and thus do the following:

- a. If the sum of the frequencies of each of the words in the sentence is greater than the value obtained from step 2, store that sentence in one file.
- b. If the sum of the frequencies of each of the words in the sentence is less than the value obtained from step 2, store those sentences in another file.

#### 4.1.1 Intuitive Justification

The summarizer must not discard the sentences from the essay but categorize them so that in the end, there is no loss of information. There are three algorithms in this summarizer that act as “selectors” for the sentences. If one algorithm misses a sentence, then the other one stores it. This ensures that none of the sentences is left out and we have categorized results. The selection algorithm then selects the most viable summary out of the three generated summaries.

## 5 The Selection Algorithm

The selection algorithm selects the best-fit summary from the list of generated summaries. The design of the selection algorithm revolves around scanning the main text and the summaries generated. The main key points that have been taken into account while designing the selection algorithm are:

- The Title generated by the Title generator should be present in the summary.
- The length of the summary should not be very long or very short. For this, the optimal length that this algorithm considers is  $1/3^{\text{rd}}$  of the original paragraph.
- The separation between the sentences is also very important. The design takes into account that the sentences selected in the summaries must not be far off from each other i.e. proximity consideration.

## 6 Title Generator

Results obtained after testing with ten documents suggest that the title generator was able to produce a meaningful title for just three of the documents and for the rest titles were not of any standard. The following key points were kept in mind while designing the title generator.

- If there is a date that occur most number of times in the document, then the title generator selected that as the title.
- If there is a proper noun like a name or a place that occur most number of times in the document, then the title generator selected that as the best title.
- If none of the above occurs, the title generator takes into account the proper noun or a date that occur in the beginning of the document.
- If there is no proper noun or a date in the document, then the title generator considers those words which occurs most number of times in the document and which are very close to each other at each occurrence.
- If none of the above works then the title generator picks up that word or words that occur most number of times in the document and which lies in the very top or beginning of the document.

## 7 Results

### 7.1 The main paragraph

*Established in 1907 by Jamshedji Nusserwanji Tata, Tata Steel is Asia's first and India's largest integrated private sector steel company. It is one of the few select steel companies that is EVA+ (Economic Value Added). Over the years, Tata Steel has emerged as a thriving, nimble, steel enterprise, due to its ability to transform itself rapidly to meet the challenges of a highly competitive global economy and commitment to become a supplier of choice by delighting its customers with services and products. Constant modernisation and introduction of state-of-the-art technology at Tata Steel has enabled it to stay ahead in the industry and successfully meet the expectations of all sections of stakeholders. Tata Steel's four-phase Modernisation Programme in the steel works has enabled it to acquire the most modern steel making facilities in the world. Its captive raw material resources and the state-of-the-art 5 MTPA (million tonne per annum) plant at Jamshedpur, in Jharkhand State, India gives it a competitive edge. Determined to be a major global steel player, Tata Steel has recently included in its fold NatSteel, Asia (2 MTPA) and Millennium Steel (now*

Tata Steel Thailand) creating a manufacturing network in eight markets in South East Asia and Pacific rim countries. Soon the Jamshedpur plant will expand its capacity from 5 MTPA to 7 MTPA by 2008. The Company plans to enhance its capacity, manifold through organic growth and investments. The Company's wire manufacturing unit in Sri Lanka is known as Lanka Special Steel, while the joint venture in Thailand for limestone mining is known as Sila Eastern. Its fifth phase of the Modernisation Programme leverages the intellectual capabilities of its employees to generate sustainable value for the stakeholders. Tata Steel is taking better Knowledge Management initiatives to shift focus from creating new physical assets to utilising them with ingenuity and a sturdy business sense.

## 7.2 Rare words display

Established in 1907 by Jamshedji Nusserwanji Tata, Tata Steel is Asia's first and India's largest integrated private sector steel company. It is one of the few select steel companies that is EVA+ (Economic Value Added). Over the years, Tata Steel has emerged as a thriving, nimble, steel enterprise, due to its ability to transform itself rapidly to meet the challenges of a highly competitive global economy and commitment to become a supplier of choice by delighting its customers with services and products. Constant modernisation and introduction of state-of-the-art technology at Tata Steel has enabled it to stay ahead in the industry and successfully meet the expectations of all sections of stakeholders. Tata Steel's four-phase Modernisation Programme in the steel works has enabled it to acquire the most modern steel making facilities in the world. Its captive raw material resources and the state-of-the-art 5 MTPA (million tonne per annum) plant at Jamshedpur, in Jharkhand State, India gives it a competitive edge. Determined to be a major global steel player, Tata Steel has recently included in its fold NatSteel, Asia (2 MTPA) and Millennium Steel (now Tata Steel Thailand) creating a manufacturing network in eight markets in South East Asia and Pacific rim countries.

## 7.3 Below average

Its captive raw material resources and the state-of-the-art 5 MTPA (million tonne per annum) plant at Jamshedpur, in Jharkhand State, India gives it a competitive edge. Soon the Jamshedpur plant will expand its capacity from 5 MTPA to 7 MTPA by 2008. The Company plans to enhance its capacity, manifold through organic growth and investments. Its fifth phase of the Modernisation Programme leverages the intellectual capabilities of its employees to generate sustainable value for the stakeholders.

## 7.4 Above average

Established in 1907 by Jamshedji Nusserwanji Tata, Tata Steel is Asia's first and India's largest integrated private sector steel company. It is one of the few select steel companies that is EVA+ (Economic Value Added). Over the years, Tata Steel has emerged as a thriving, nimble, steel enterprise, due to its ability to transform itself rapidly to meet the challenges of a highly competitive global economy and commitment to become a supplier of choice by delighting its customers with services and products. Constant modernisation and introduction of state-of-the-art technology at Tata Steel has enabled it to stay ahead in the industry and successfully meet the expectations of all sections of stakeholders. Tata Steel's

*four-phase Modernisation Programme in the steel works has enabled it to acquire the most modern steel making facilities in the world. Determined to be a major global steel player, Tata Steel has recently included in its fold NatSteel, Asia (2 MTPA) and Millennium Steel (now Tata Steel Thailand) creating a manufacturing network in eight markets in South East Asia and Pacific rim countries. The Company's wire manufacturing unit in Sri Lanka is known as Lanka Special Steel, while the joint venture in Thailand for limestone mining is known as Sila Eastern. Tata Steel is taking better Knowledge Management initiatives to shift focus from creating new physical assets to utilising them with ingenuity and a sturdy business sense.*

### **7.5 Title**

Established 1907 Jamshedji Nusserwanji Tata

### **7.6 The Final Selected Summary**

The summary with rare words.

## **8 Conclusion**

The algorithms illustrated can be regarded to be among the simplest of the text summarization algorithms. However, the results that they produced are worth to be noted. The main aim of the research was to understand as to what humans think when choosing a particular word while framing the sentences. One important focus was to summarize small pieces of texts rather than larger ones as books. The main ingredient of this summarization system is the selection algorithm. The selection algorithm is the main core of the algorithm, which finally gives to the user the best-summarized content. An automatic text summarizer has to satisfy the reader in the end and that remains the main goal.

## **Acknowledgement**

We thank the Department of Computer Science and Engineering for providing the computing facilities and resources.

## **References**

1. Eduard H., *Text Summarization*. Chapter 32.
2. Berger A.L., Mittal V.O.: OCELOT: A System for Summarizing Web Pages, Proceeding of the 23rd Annual International ACM SIGIR Conference on Research and

- Development in Information Retrieval, p. 144–151, July 24–28, 2000, Athens, Greece (2000)
- 3. Buyukkokten O., Garcia-Molina H., Paepcke A. Seeing the while in parts: Text Summarization for Web Browsing on Handheld Devices, Proceedings of the 10th International Conference on World Wide Web, p. 652–662, May 01–05, 2001, Hong Kong (2001)
  - 4. Hu, Y., Xin, G., Song, R., Hu, G et al... Title Extraction from Bodies of HTML Documents and its Application to Web Page Retrieval. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2005)
  - 5. Xue Y., Hu Y., Xin, G., Song, R., Shi S., et al. : Web Page Title Extraction and its application, Information Processing and Management: An International Journal .43, 1332–1347, September, 2007.
  - 6. Shoaib J.M. et al. : Enhancements in Query Evaluation and Page Summarization of The Thinking Algorithm. In the Proceedings of the Third International Symposium on Information Technology, Kuala Lumpur, Malaysia, vol. III, pp. 1979–1987
  - 7. Lin C-Y, Hovy E. : Identify Topics by Position. Proceedings of the 5th Conference on Applied Natural Language Processing, March(1997).
  - 8. Dalianis H., Hovy. E.: Aggregation in Natural Language Generation. In Adorni, G. & Zock, M. (Eds.), Trends in Natural Language Generation: an Artificial Intelligence Perspective, EWNLG'93, Fourth European Workshop, Lecture Notes in Artificial Intelligence, No. 1036, pp. 88–105, Springer Verlag (1996)
  - 9. Dalianis H.: ASTROGEN – Aggregated deep and Surface naTuRal language GENerator [Online] Available:<http://www.dsv.su.se/~hercules/ASTROGEN/ASTROGEN.html> (1999)
  - 10. Marcu D. :From Discourse Structures to Text Summaries. The Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization, pp 82–88, Madrid, Spain.,July (1997)
  - 11. McKeown K., Radev D. :Generating summaries of multiple news articles. In Proceedings, 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 74–82, Seattle, Washington, July (1995)
  - 12. Luhn, H.P., Automatic Creation of Literature Abstracts. IBM Journal (1958), 159–165.
  - 13. Edmundson, H.P., New Methods in Automatic Extracting, Journal of the ACM 16(2): 264–285 (1958)
  - 14. Hassel, M., Dalianis H. : Generation of Reference Summaries. In the proceedings of the 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, April 21–23 2005, Poznan, Poland.
  - 15. Chuang, W.T., Yang, J. : Text Summarization by Sentences Segment Extraction Using Machine Learning Algorithms. Springer Berlin/Herdelberg 454–457. July 13, (2007)
  - 16. Wang, J., Zhou, S., Hu, Yun-Fa. : Sentence Clustering based automatic Summarization. Machine Learning and Cybernetics, 2003 International Conference on Digital Object Identifier 1, 57–62 (2003).
  - 17. Mitra, M., Singhal, A., Buckley, C. : Automatic Text Summarization by Paragraph Extraction. In Proceedings of the ACL '97/EACL '97 Workshop on Intelligent Scalable Text Summarization (Madrid, Spain, 1997), pp 31–36 (1997)

# A Hybrid Approach for Transliteration of Name Entities

R.C.Balabantaray<sup>1</sup>, S.Mohanty<sup>2</sup> and R.K. Das<sup>3</sup>

<sup>1</sup>International Institute of Information Technology  
Bhubaneswar, Orissa, India

Email: rakesh\_b\_ray@yahoo.co.in

<sup>2,3</sup>Dept. of Computer Science & Application  
Utkal University, Bhubaneswar,Orissa,India

Email: sangham1@rediffmail.com, dasranjan257@yahoo.co.in

**Abstract:** To develop a system for translation of one language to another is one of the most important research challenges in Artificial Intelligence (AI). In Machine Translation (MT) the name entity recognition (NER) is one of the most challenging task. In this paper we propose a new statistical method for transliterating the identified name entities based on the linguistic knowledge of possible conjuncts and diphthongs in source and target language . The work presented in this paper is part of a larger effort to develop MT system which can take care of name entities.

## 1 Introduction

Name Entity Recognition (NER) is an information extraction task that is concerned with the recognition and classification of name entity from free text [Grishman, R., 1997]. Name entities classes are, for instance, locations, person names, organization names, dates, times and money amounts. For example, in the sentence: "*The Chief Minister of Orissa Mr. Naveen Patnaik had delivered a talk on cultural issues in a function organized at Bhubaneswar.*", a name entity recognition process looking for named persons and locations would identify the person name Naveen Patnaik and the location Bhubaneswar. This recognition can be based on a variety of features of the terms, the sentence, the text and its syntax and could leverage external sources of information such as thesauri and dictionaries, for instance. In the example, a system may have applied a simple rule guessing that the capitalize words directly following the term '*Minister*' are name of a person. But the most important question is how to convert these name entities to appropriate words in the target language so, that the Machine translation will be fruitful.

Named entity recognition (NER), the identification of entity names in free text, is a well-studied problem. In most previous work, NER has been applied to news articles e.g., (Bikel et al., 1999; McCallum and Li, 2003), scientific articles e.g., (Craven and Kumlien, 1999; Bunescu and Mooney, 2004)), or web pages e.g., (Freitag, 1998).

In Natural Language Processing (NLP) application areas such as information retrieval, question answering systems and machine translation, there is an increasing need to translate out of vocabulary (OOV) words from one language to another. They are translated through transliteration, the method of translating into another language by expressing the original foreign words using characters of the target language preserving the pronunciation in their original languages. Thus, the central problem in transliteration is predicting the pronunciation of the original word. Transliteration between two languages, that use the same set of alphabets, is trivial: the word is left as it is. However, for languages that use different alphabet sets, the names must be transliterated or rendered in the target language alphabets. Technical terms and named entities make up the bulk of these OOV words. Named entities hold a very important place in NLP applications. Proper identification, classification and translation of named entities are very crucial in many NLP applications and pose a very big challenge to NLP researchers. Named entities are usually not found in bilingual dictionaries and they are very productive in nature. Translation of named entities is a tricky task: it involves both translation and transliteration. Transliteration is commonly used for named entities, even when the words could be translated. Different types of named entities are translated differently. Numerical and temporal expressions typically use a limited set of vocabulary words (e.g., names of months, days of the week etc.) and can be translated fairly easily using simple translation patterns. The named entity machine transliteration algorithms presented in this work focus on person names, locations and organizations. A machine transliteration system that is trained on person names is very important in a multilingual country like India where large name collections like census data, electoral roll and railway reservation information must be available to multilingual citizens of the country in their vernacular. In the present work, the proposed model has been evaluated on a training corpus of person names.

India is the second largest in population in the world with more than one billion populations. There are 19 constitutional languages with 10 scripts and over 1650 dialects. Orissa is a state of India situated in the eastern region, with a population of 36.7 million according to 2001 census. Orissa is the first state in India to have been formed on linguistic basis. Oriya is the official as well as spoken language of Orissa, and is one of the constitutional languages of India. We are working towards developing a robust system (OMTrans) which will translate the source language English to target language Oriya. In this OMTrans system the transliteration of name entities will be crucial for effective translation.

## 2 Related Work

A hybrid neural network and knowledge-based system to generate multiple English spellings for Arabic personal names is described in (Arbabi et al., 1994). (Knight and Graehl, 1998) developed a phoneme-based statistical model using finite state transducer that implements transformation rules to do back-transliteration. (Stalls and Knight, 1998) adapted this approach for back transliteration from Arabic to English for English names. A spelling-based model is described in (Al-Onaizan and Knight, 2002a; Al-Onaizan and Knight, 2002c) that directly maps English letter sequences into Arabic letter sequences with associated probability that are trained on a small English/Arabic name list without the need for English pronunciations. The phonetics-based and spelling-based models have been linearly combined into a single transliteration model in (Al-Onaizan and Knight, 2002b) for transliteration of Arabic named entities into English. Several phoneme-based techniques have been proposed in the recent past for machine transliteration using transformation-based learning algorithm (Meng et al., 2001; Jung et al., 2000; Vigra and Khudanpur, 2003). (Abduljaleel and Larkey, 2003) have presented a simple statistical technique to train an English-Arabic transliteration model from pairs of names. The two-stage training procedure as described in (Abduljaleel and Larkey, 2003), “first learns which n-gram segments should be added to unigram inventory for the source language, and then a second stage learns the translation model over this inventory”. This technique requires no heuristic or linguistic knowledge of either language. (Goto et al., 2003) described an English-Japanese transliteration method in which an English word is divided into conversion units that are partial English character strings in an English word and each English conversion unit is converted into a partial Japanese Katakana character string. It calculates the likelihood of a particular choice of letters of chunking into English conversion units for an English word by linking them to Katakana characters using syllables. Thus the English conversion units consider phonetic aspects. It considers the English and Japanese contextual information simultaneously to calculate the plausibility of conversion from each English conversion unit to various Japanese conversion units using a single probability model based on the maximum entropy method. (Haizhou et al., 2004) presented a framework that allows direct orthographical mapping between English and Chinese through a joint source-channel model, called n-gram transliteration model. The orthographic alignment process is automated using the maximum likelihood approach, through the Expectation Maximization algorithm to derive aligned transliteration units from a bilingual dictionary. The joint source-channel model tries to capture how source and target names can be generated simultaneously, i.e., the context information in both the source and the target sides are taken into account. A tuple n-gram transliteration model (Marino et al., 2005; Crego et al., 2005) has been loglinearly combined with feature functions to develop a statistical machine translation system for Spanish-to-English and English-to-Spanish translation tasks. The model approximates the joint probability between source and target languages by using trigrams. The present work differs from (Goto et al., 2003; Haizhou et al., 2004) in

the sense that identification of the transliteration units in the source language is done using regular expressions and no probabilistic model is used. The proposed modified joint source-channel model is similar to the model proposed by (Goto et al., 2003) but it differs in the way the transliteration units and the contextual information are defined in the present work. No linguistic knowledge is used in (Goto et al., 2003; Haizhou et al., 2004) whereas the present work uses linguistic knowledge in the form of possible conjuncts and diphthongs in Bengali.

### 3 Methodology

A transliteration system takes as input a character string in the source language and generates a character string in the target language as output. The process can be conceptualized as two levels of decoding: segmentation of the source string into transliteration units (TU); and relating the source language transliteration units with units in the target language, by resolving different combinations of alignments and unit mappings. The problem of machine transliteration has been studied extensively in different paradigm. For a given English name E , we have to find out the most likely Oriya transliteration O that maximizes  $P(O|E)$ . Applying Bayes' rule, it means to find O to maximize

$$P(E,O) = P(E|O) * P(O) \quad (1)$$

with equivalent effect. This is equivalent to modelling two probability distributions:  $P(O|E)$ , the probability of transliterating O to E , which is also called transformation rules, and  $P(O)$ , the probability distribution of source, which reflects what is considered good English transliteration in general. Likewise, in Oriya to English transliteration, we could find E that maximizes

$$P(E,O) = P(O|E) * P(E) \quad (2)$$

for a given Oriya name. In equations (1) and (2),  $P(O)$  and  $P(E)$  are usually estimated using ngram language models.

The English transliteration units are of the form C\*V\* where C represents a consonant and V represents a vowel whereas the transliteration units in Oriya words take the pattern C<sup>+</sup>M where C represents a vowel or a consonant or a conjunct and M represents the vowel modifier or matra.

#### *Algorithm*

Step1: The phoneme grouping (pseudo syllable forming) will be done.

Step 1.1: The string will be scanned from left to right once we will find a vowel the preceding consonants will be grouped into a conjunct group (CG/TU) with vowel becoming Matra. (In exception if 2 vowels appear in conjunction then each will be treated as separate group or syllable)

Step 2: Mapping of the English CG/TU will be done to its equivalent Oriya CG/TU. A lot of possible outcomes will come out in this step. (refer to Table 1)

Step 3: context sensitive rules will be applied to restrict the possibilities based on monogram, bigram and trigram models upon the CG/TUs. (refer to Table 2)

Step 4: The checking will be done with the Name database to figure out the correct transliterated output based on statistical augmenting techniques.

Step 4.1: the string of TUs will be divided into two parts called as PRE (initial 2-3 TUs) and POST(last 2-3 TUs of the string as normally maximum size of string will be five in most cases and six in very rare cases).

Case 1: 2 TUs or 3TUs or check the whole word

Case 2: match from 1<sup>st</sup> to right when no matching then POST= unmatched TU to end.

Step 4.2: check POST from the end to the beginning

if matched map Eng to corresponding Oriya TUs.

Else check in the Eng-Oriya table for the higher probability of occurrence (based on intuition and frequency count).

**Table 1.** Shows the various possible mapping of the English alphabet set to corresponding Oriya character set.

English Alphabets	Equivalent Oriya characters
a	ା / ଅ (some times matra)
c	କୁ / କ୍ଷ
d	ଙ୍କ / ଖ
i	ି / ଇ (matra)
l	ିମ୍ବ / ମ୍ବ
n	ନ୍ତର୍ମୁ / ଞ୍ଚ
s	ଜ୍ବ / ହ୍ର / ଶ୍ର
t	ତ୍ର୍ପୁ / ପ୍ର
u	ଦ୍ଵୀପୁ / ଏ (matra)

**Table 2** Shows the various possible combination of Oriya characters in different context

Rule Number	Allowed Combinations in a TU of Oriya language
1	nk(h) - * / <
2	ng(h) - = / +
3	nt(h) - ñ / x / Ä / ¶
4	nd(h) - t / u / .. / -
5	ks - µ
6	ksh - l
7	sk - ^2 / '
8	st(h) - j ñ / y / ½
9	sch - ¾
10	shna - »
11	sp - È
12	ps - Ð
13	ntr - § / ÄÖ
14	tsn - j úá
15	ndr - t ö

#### 4 English to Oriya Machine Translation

Let us take the word “**chandrakanta**” as an example to test our system. This word contains 12 characters with the possible number of interpretation of them as follows.

Eng	c	h	a	n	d	r	a	k	a	n	t	a
Oriya	ଛ	ହ	ଅ	ନ୍ତର	ଦା	ର	ଅ	କନ୍ତା	ଅ	ନ୍ତର	ତ	ଅ
possibility	2	1	2	2	2	1	2	1	2	2	2	2

All together this is going to generate  $2^9$  possibilities.

In this word there are 3 types of consonants characters.

- (i) two consonant forming one consonant character.

ch: Q/R

- (ii) two consonants forming one conjunct character

nt: \_ñ / Ā

- (iii) three consonants forming one conjunct character.

ndr: tö / ..ö

Step 1:

Following the basic combination pattern of consonants characters as a unit we get 8 units.

Eng	ch	a	ndr	a	k	a	nt	a
Oriya	Q R	ଅ ା	tö ..ö	ଅ ା	L ା	ଅ ା	ନ୍ତ ା	ଅ ା
possibility	2	2	2	2	1	2	2	2

Step 2:

Following the principle of homographic articulation this step restricts the (nasal+stop) sound combinations like nd, nt to be interpreted as \_ + ] / Z + X and \_ + [ / Z + V respectively and exclude the possibilities of \_ + X / Z + ] / \_ + V / Z + [. Then we get  $2^7$  possibilities.

Step 3:

Forming of TUs.

Input: chandrakanta

Output: cha\_ndra\_ka\_nta

So, there are 4 TUs and we will have 4 graphemes(characters) in Oriya.

Total possible outcomes for the whole word is  $4*4*2*2=2^7$  numbers. Our task is to sort all other except exactly one, the correct /desired one.

#### Step 4:

Checking with the name database in Oriya.

##### Step 4.1:

Break the string of TUs into PRE+POST. Total 4 TUs break into 2+2

Cha\_ndra\_ka\_nta => cha\_ndra+ka\_nta

Check the PRE TUs with the database from left the possibilities are Qtö, Qtö, QtöD, QDtö, Rtö, RDtö, RtöD, RDtö. All total it is 8 in numbers ( $2^3$ ).

All other except Qtö and QtöD are invalid in Oriya and so won't match. again amongst these two Qtö will match in almost all cases as QtöD is very very rare.(but possible in feminine naming)

##### Step 4.2:

Checking POST from right. 2 possable matching are LĐ\_ñ, LÄĐ out of 8 possibilities. Higher percentage of matching will be with LĐ\_ñ (>98) and LÄĐ vey rarely. Finally by joining PRE and POST we get QtöLĐ\_ñ.

## 5 Conclusion

Till today few works has been done in the field of transliteration of name entities from English to Indian language. But in the field of transliteration from Indian language to English some work has been done (for the name sake Bengali to English) which involves less complexity. We have made a novel attempt in this paper for solving this complex task which can also be applicable to other Indian languages. The test result of our approach is quiet impressive. We have tested our algorithm on a sample set of 1000 standard names and we getting accuracy in the range of 65–70%.

## References

1. Nasreen, A.J., Larkey, L.S.,:Statistical Transliteration for English-Arabic Cross Language Information Retrieval. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM 2003), New Orleans, USA, 139–146.(2003)
2. Al-Onaizan Y. and Knight K.: Named Entity Translation: Extended Abstract. Proceedings of the Human Language Technology Conference (HLT 2002), 122–124 (2002)

3. Al-Onaizan, Y. and Knight, K.: Translating Named Entities Using Monolingual and Bilingual Resources. Proceedings of the 40th Annual Meeting of the ACL (ACL 2002), 400–408 (2002).
4. Al-Onaizan, Y., Knight, K.: Machine Transliteration of Names in Arabic Text. Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages (2002)
5. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In CoNLL (2003)
6. Arbab, M., Scott, M., Fischthal, V., Cheng C., Bar E.: Algorithms for Arabic name transliteration. IBM Journal of Research and Development, 38(2), 183–193. (1994)
7. Ekbal, A., Naskar, S.K., Bandopadhyaya, S.: A modified joint source-channel model for transliteration. Proceedings of the COLING/ACL on Main conference poster sessions. Sydney, Australia, pp: 191–198 (2006)
8. Crego J.M., Marino J.B., de Gispert, A.: Reordered Search and Tuple Unfolding for Ngrambased SMT. Proceedings of the MT-Summit X, Phuket, Thailand, 283–289 (2005)
9. Bikel D. M., Schwartz R. L., Weischedel R. M.: An algorithm that learns what's in a name. Machine Learning, 34, 211–231 (1999)
10. Freitag, D: Information extraction from html: application of a general machine learning approach. In AAAI–98 (1998)
11. Goto I., Kato, N., Uratani, N., Ehara. T.: Transliteration considering Context Information based on the Maximum Entropy Method. Proceeding of the MT-Summit IX, New Orleans, USA, 125–132 (2003)
12. Grishman, R.: Information Extraction: Techniques and Challenges”, Lecture Notes in Computer Science, Vol. 1299, Springer–Verlag (1997)
13. Li, H., Min., Z. Jian, S.: A Joint Source–Channel Model for Machine Transliteration. Proceedings of the 42nd Annual Meeting of the ACL (ACL 2004), Barcelona, Spain, 159–166 (2004)
14. Young, J. S., , Hong S. L., Paek E., An English to Korean Transliteration Model of Extended Markov Window. Proceedings of COLING 2000, 1, 383–389 (2000)
15. Knight K., Graehl, J.: Machine Transliteration, Computational Linguistics, 24(4), 599–612 (1998)
16. Marino J. B., Banchs R., Crego J. M., A. de Gispert, P. Lambert, J. A. Fonollosa and M. Ruiz, Bilingual N-gram Statistical Machine Translation. Proceedings of the MT–Summit X, Phuket, Thailand, 275–282.
17. Craven M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. In ISMB–99 (1999)
18. Meng Helen M., Wai–Kit Lo, Chen, B., Tang. K.: Generating Phonetic Cognates to handle Name Entities in English–Chinese Crosslanguage Spoken Document Retrieval. Proceedings of the Automatic Speech Recognition and Understanding (ASRU) Workshop, Trento, Italy (2001)

19. Bender, O., Josef Och F., Ney, H.: Maximum Entropy Models for Named Entity Recognition In: Proceedings of CoNLL–2003, Edmonton, Canada, pp. 148–151 (2003)
20. Bunescu R., Mooney. R. J.: Relational markov networks for collective information extraction. In ICML–2004 Workshop on Statistical Relational Learning (2004)
21. Stalls, B.G., Knight K.: Translating names and technical terms in Arabic text. Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages, Montreal, Canada, 34–41. (1998)
22. Paola, V., Khudanpur, S.. Transliteration of Proper Names in Crosslingual Information Retrieval. Proceedings of the ACL 2003 Workshop on Multilingual and Mixedlanguage Named Entity Recognition, Sapporo, Japan, pp.57–60. (2003)
23. Mohanty, S., Balabantaray, R.C.: Name Entity Recognition in OMTrans. Communicated to International Journal of Translation (2007)

# **Information Uptriever: A system for Content Assimilation and Aggregation for Developing Regions**

Ravindra Dastikop, G. A. Radher and Jaydevi C. Karur

Department of Computer Science and Engineering  
SDM College of Engineering & Technology  
Dhavalgiri, Dharwad -580002, INDIA.

ravindra.dastikop@gmail.com, radderga@gmail.com, jck1965@gmail.com

**Abstract:** In the developing regions much of local information content is not available in forms easily indexable by web crawlers. In this paper we describe a new local data uploading system called *Information Uptriever*. The goal of Information Uptriever is to selectively upload local data relevant to a pre-defined set of decision-making situations. Rather than uploading normal (or all) information that does not necessarily result in an immediate action an *Information Uptriever* uploads actionable information that is likely to be most relevant in a decision-making situation and may lead to action. Using freely available resources on the web we implement the system for uploading actionable information pertaining to the domain of Indian higher education as a proof-of-concept. Our anecdotes suggest that uptrieving is very effective for uploading information relevant to a pre-defined set of actions. The live system is accessible at <http://uptriever.googlepages.com>

## **1 Introduction**

The scarcity of local information content on the WWW prevailing in the developing regions motivates the work described in this paper. Researchers recently have made efforts on designing systems, both in concept and implementation, for extending the scope of WWW beyond the connectivity accomplished developed world to the counties in the developing regions. Several works have addressed the issues of how to overcome the access barriers, low-connectivity and local content challenges prevailing in these countries [1,2,3,4]. The information needs of these people are very often local. They are interested in knowing the educational course announcements, job openings in their region. This kind of information is not globally required and hence is typically not currently available on the web. More importantly, hardly any channels exist for these people

to contribute such information for others (consumers living within the regions and elsewhere) consumption. Therefore, it is important to find ways for this population (the producers and consumers of local information) to leverage now available web resources and suitably adapt the World Wide Web to their own needs. In this paper we address the challenge of local information content assimilation and aggregation. We describe a new local data uploading system called an *Information Uptriever*. The goal of an *Information Uptriever* is to selectively upload local data that are relevant to a pre-defined set of decision-making situations. The situations are pre-identified and likely to be faced by a sizeable number of users. Rather than uploading all available general information that might not lead into action an uptriever uploads data that are likely to be most relevant for the decision-making situation and avoids irrelevant data uploading. This leads to significant savings in uploading efforts, and helps keep the uploaded data more up-to-date for those selected decision-making situations.

## 2 Motivational Examples

In the developing countries most of the consumers and producers of information, goods and services are relegated to local markets in geographical vicinity. Some of the examples include local newspapers (information) and educational institutes.

Here are some motivational examples

1. Student seeking information about a new or existing course in another city or region may not able to get this because institutions usually publish them in local newspapers.
2. A faculty specialized in a particular stream contemplating a career move may be interested in the institutions adding a new course in her specialization or new institution coming up anywhere in the country.
3. An NRI planning a repatriation in coming years may be interested in such information for planning for her children and wards
4. Educational seekers in neighboring developing countries may be interested in such information

Today such information is not easily available for wider audiences for various reasons. Except for top few, most institutions do not put such information on their web sites and many institution have home-page only kind of web sites. In order to reach wider markets, it is important for this local information to reach wider audiences. Although we have used motivational examples from the domain of education, examples can be drawn from any other domain as well (patients/professionals seeking information on specialty hospitals)

### 3 Our Proposed Solution

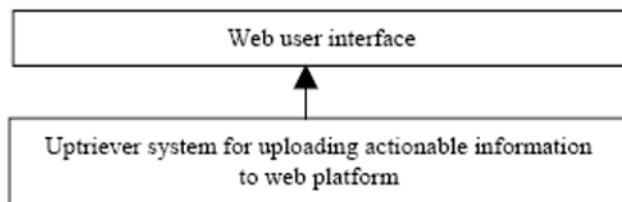
In this paper we demonstrate a new local data uploading system called *Information Uptriever*. The goal of Information Uptriever is to selectively upload local data relevant to a pre-defined set of decision-making situations.

**Definition:** An Information Uptriever system is designed for identifying, acquiring, assimilating, aggregating and uploading global and local information that is relevant for developing regions to the WWW platform that improves the reach of the information.

*Information Uptriever* draws upon a number of core ideas, concepts and technologies in its design and implementation. One of these key concepts is Information Uptrieval [3] that articulated the need for such system. The concept of actionable information [4] provides basic unit of information to be considered for Uptrieval process. One of key technologies the uptriever uses is a suite of online publishing tools that provide resources for uploading and publishing content onto web platform. Their functionality such as API plays key role in converting otherwise non-crawable data sources into digital form and then makes it searchable by search engines. Information uptriever system architecture

#### 3.1 Uptriever system architecture

Figure 1 illustrates a simplified diagram of the system architecture. The power of Uptriever lies in variety of data sources it can cover.



**Fig. 1.** Uptriever system architecture

#### 3.2 Uptriever algorithm

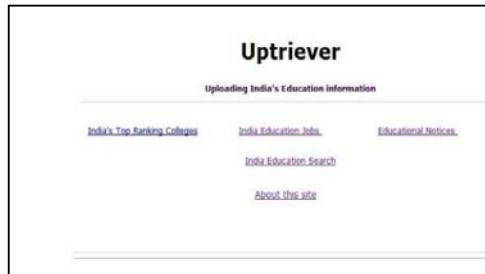
Uptriever exploits a set of freely available online publishing tools such as blogs, docs, etc to publish actionable information such as course announcements, examination dates on to web platform.

- Information Uptriever algorithm**
1. Accept file to be uploaded
  2. Choose and invoke uploading service and pass the file named in step 1.
  3. Obtain the URL of file uploaded
  4. Update URL file (add/append new URL)
  5. Publish the file at web user interface
  6. Repeat step 1-5

**Fig. 2.** Information Uptriever Algorithm for uploading local data to the WWW platform

#### 4 Prototype and Implementation

To demonstrate the idea of Information Uptriever we have developed a prototype. Although, the Information Uptriever can be used for uploading any kind of information, for a proof-of-concept prototype we have used the system for uploading various type of higher education information. A working prototype is available for live demo at <http://uptrivever.googlepages.com>



**Fig. 3.** Web user interface for Uptrieved Information

#### 5 Experimental Evaluation

*Information Uptriever* system is implemented to publish local data (pertaining to Indian higher education) to the web at <http://uptrivever.googlepages.com/more>. We are promoting it through email communications. We evaluate system by way of monitoring visitor data through freely available analytics tools and results so far are found to be encouraging.

## 6 Conclusions and Future Work

In this paper, we demonstrated a novel system called “Information Uptriever”, which allows people in developing regions to identify and upload local data to web platform and thus allows them to contribute such information for others’ consumption. The consumers of such local data may be just miles away from each other or in some instances many be separated by continents.

**Contributions:** This approach presents a number of advantages:

- It can be applied to any identified decision-making situation
- It can be built using existing freely accessible resources on the web
- It can be combined with web information retrieval systems and presented as aggregated search system such as <http://edusense.googlepages.com>

**Future work:** In future we plan to strengthen our system with the use of sophisticated tools such as UIMA and RIA frameworks. Our next step is to automate the Uptrieval process with the use of publicly available APIs..

## References

1. Thies W., et al :Searching the World Wide Web in Low-Connectivity Communities, 7-11 May 2002 WWW2002, Honolulu, Hawaii, USA (2002)
2. Web Delivery Models, Panel Discussion WWW 2007, May 8-12, 2007, Banff, Canada (2007)
3. Agarwal S., Kumar A., Mukherjea S., Nanavati A., Rajput N.: Information Uptrieval: Exploring Models for Content Assimilation and Aggregation for Developing Regions, WWW 2008 / Panel Overview April 21-25, 2008 · Beijing, China (2008)
4. Talukder A.: Web-Greatest Equalizer for the Developing World, WWW 2007, May 8-12, 2007, Banff, Canada (2007)

# Ranking of Products through Blog Analysis

Niladri Chatterjee<sup>1</sup>, Sumit Bisai<sup>1</sup> and Prasenjit Chakraborty <sup>2</sup>

<sup>1</sup>Department of Mathematics, IIT, Hauz Khas, New Delhi – 110016  
niladri@maths.iitd.ac.in, sumitbisai@gmail.com,

<sup>2</sup>IBM India Pvt Ltd, Bangalore- 560071  
cprasenj@in.ibm.com

**Abstract:** Many web blogs contain comments of users on different products. Such comments are often helpful for a naive user to decide which particular product (from among various alternatives available in the market) he/she is interested in. However, manual analysis of these blogs is time-consuming. In this work we propose a tool for automatic analysis of blogs on different products, and rank them on the basis of certain key features. The task however is not straightforward as users' comments are often replete with ungrammatical or poorly -structured sentences, incoherence of themes, usages of synonyms, and many other usual NLP problems, which need to be dealt with appropriately. The overall procedure takes the following steps: Preprocessing, POS tagging, Brand identification, Feature identification and Ranking. The scheme has been tested on blogs of Laptops and Automobiles. The initial results are found to be very promising.

## 1 Introduction

World Wide Web has revolutionized the way we gather information today with plentiful of web sites giving information of different kinds. Such a large-scale availability of information on one hand extends the gamut of applications of web-based technology. On the other hand, it necessitates the development of newer techniques for effective and efficient utilization of the stored knowledge.

In this work we look at a possible application of web-based knowledge in classification of products based on users' comments on them. For example, suppose a person wants to purchase a laptop. The market is now flooded with a large varieties of laptops with wide range of prices, and wide choice of features, such as Wi-Fi with a good range, multimedia friendliness, fingerprint reader. A person may be interested in some of these features, and/or may dislike some of them. Worse still, the buyer may not even have clear idea of the availability of features, as changes happen pretty fast in these days. Hence making up one's mind is not easy.

A practical solution may lie in reviewing users' comments on their purchases. There are sites where people share their opinions on various products of daily use.

These are typically called “blogs” (derived from “web logs”). Web blogs provide commentary or news on a particular subject. The ability for readers to leave comments in an interactive format is an important part of many blogs.

There are blogs where people share their opinions of different services and products. A person desirous of purchasing a laptop may refer to such blogs and glean opinions of different users on the pros and cons of different models available in the market. This allows the potential buyer to gain insight into the availability of features, prices, and also the positive and negative aspects of each model. As an example, consider the case of a user who wants to browse through all products which adhere to his specification and then make a decision. His specification could be as abstract as just being “good” or “excellent”, or can be precisely given within some range, e.g. cars below Rs. 5 Lakhs. With this requirement no search engine today can serve his purpose to display only those which matches his choice. Typically the user has to read all the pages and then sort out things, which is definitely time-consuming. A further problem is that the information/knowledge stored in this form is often much unstructured: a typical blog combines text, images, and links to other blogs, Web pages, and other media related to its topic. Moreover, as users express their views in natural languages (we concentrate on English only) the expressions suffer from certain drawbacks. For illustration consider the following.

**Incoherence of topic:** *The French have seven types of love. Eskimos have 40 words for snow. Jews have 78 ways to call you the village idiot. As such, pistonheads need a few ways to explain “ugly.” The Pontiac G8’s face is without question Exciting Ugly.* [1]

**Ungrammatical sentences:** *u dont know bout the fun wid the big engine nd muscular body...dont go wid face of Ambassador. Press the full Xelerator in IIInd gear for 7 second and watch what happns.* [2]

Further, common NLP problems, such as, synonymous words, ambiguity of word senses, anaphora [3] do exist in the reviews. As a consequence, specific tools need to be developed to achieve the desired goal.

Our work aims at developing such a tool. The information in its entirety is so huge that to develop an effective technique for extraction of useful sentences for all domains is a distant far. But when the technique is confined to a particular domain or set of domains we may discover some pattern through which we can extract the meaningful information. In this work we have developed such a technique for extraction of useful sentences (based on its features) for a particular product from the comments that reviewers have posted in different web-blogs.. We then used the extracted information to rank the products. Because of space restriction we limit our discussions to two different categories of products: laptops and automobiles, and the efficacy of the method is shown.

## 2 Method for Web Information Extraction and Classification: The Proposed Approach

From the manual analysis of the reviews posted in the blogs we inferred that most of the sentences describe some feature or sub-part of a product. However, there are quite a good number of sentences which do not specify any feature but describe the product as a whole. The rest of the sentences are too vague to extract any meaning from them. Of the useful sentences describing either a feature or the product we found that usually the first sentence refers to the feature explicitly, whereas the subsequent sentences refer to it implicitly. For illustration, consider the following [4]: *The X301 employs an LED-backlit, 13.3-inch (1440 x 900-pixel resolution) display with a matte finish that prevents most glares. That high resolution certainly lets you see more of documents and Web pages, though the default text may be too small for some eyes.*

But this does not always hold true. Many a time there are situations where a sentence neither contains any feature, nor does it refer to the feature discussed in the previous sentence. For example, consider the following pair of sentences [5]: *Mind you, the LX is no sports sedan. There is neither the power underfoot nor the chassis control needed for genuine hustling.*

Among the sentences which describe some feature we found that many do specify numerically the technical specification for that feature e.g. diameter of the wheel is 18 inches. Such sentences we have ignored for most of the cases as the user can very easily refer the *technical specification sheet* (TSS) for that particular product. Instead we have focused on the sentences which either enumerate the merit or demerit of any feature or the product. This leaves us with extracting only those sentences which have the intersection of describing a feature or the product and also which uses some adjective or adverb to classify it.

Extraction of useful sentences describing some feature is another major task. For this we need to identify the important features a product can have based on which users are likely to choose some model/brand over another. Most of the features can be easily enumerated by looking at the TSS for that product, and hence for other products of the same category also<sup>1</sup>. However, after going through different user reviews we found that often the same feature is expressed under different names creating confusion in the user's mind. For example in a blog on laptops one may find different synonyms like 'notebook', 'lifebook', 'system', 'lapi' etc. each essentially describing the 'laptop'. Hence we needed to have a ready list of synonyms for each such feature. We have created a database of commonly used terms for storing (and subsequent retrieval) of a feature and its possible synonyms.

---

<sup>1</sup> TSS is often available from the relevant site of the product under consideration. E.g. [http://www1.ap.dell.com/content/products/features.aspx/featured\\_notebook6?c=in&cs=indhs1&l=en&s=dhs](http://www1.ap.dell.com/content/products/features.aspx/featured_notebook6?c=in&cs=indhs1&l=en&s=dhs) provides the technical specifications of Dell XPS M1530 laptop.

### 3 Methodology and Implementation Details

**PreProcessing:** To overcome the diversity in expression of user reviews from different geographical regions we have pre-processed the figurative, decorative and slang sentences by mutating them to standard one without losing their meaning and complexity.

**POS Tagger:** The first level filtration applied over the whole pre-processed blog is to identify the sentences having some adjective or adverb. We run the POS tagger through all the sentences and extract sentences having some adjective or adverb in them, as these sentences typically contain the users' comments on features.

**Brand Identification:** A blog may refer to multiple products and models for mutual comparison. Hence the task of identifying which brand the blog actually refers to becomes significant. We maintain a list of companies offering the product in that category. From the analysis of the blogs we found that the model name typically comes after the company name e.g. "... Sony VAIO T2310...". Hence one can search for a match of the company name from our database and can pick the next K words as a possible name of the model. We found that the values of K lie generally between 2 and 3 for most of the product types. The algorithm starts looking at these K words one by one and if it finds the word to be a noun we include it in the name of the brand. Once either we find a word which is not a noun or we reach the end of the K words, we stop the brand name formation. Applying this technique over the whole blog we have a set of such K or less word elements. As a brand name we pick the one that has the highest frequency of occurrence and take its largest number of words present in the set. This identifies the particular brand the blog is talking about when it is compared against other products or brands.

**Feature Identification:** Once we have the list of useful sentences we process each sentence and identify the relevant feature. A sentence may talk about only a single feature, or many features. We call these two types of sentences *singleton* and *collecton*, *respectively*. For identification of a sentence as singleton or collecton we match the features that are already picked (and stored) from the technical specification sheet (TSS) and count the number of matches for each entry from the sheet. However, we need to consider not only the entry from the TSS but also the possible common synonyms that are used to describe the feature. Due to the presence of large number of synonyms we normalize each feature and keep a base name for it which is often same as the technical name. Any appearance of a synonym only contributes to the local count of that word. But it doesn't affect any increase in the global count, which keeps a count of the total number of distinct features in the sentence. Note that global count helps to categorize a sentence as singleton or collecton. The local count, as seen later, helps us in analyzing the sentence precisely.

We found that many sentences talk about the overall product instead of any feature. Hence all possible synonyms which may refer to that category of product should contribute towards the overall qualitative analysis of the product. The simplest case for feature identification is with the singleton sentences where we

have already identified the feature it is referring to. A collecton, on the other hand, may fall in two categories where it may be a singleton but due to the presence of the reference of overall product (e.g. laptop, car) we have put them otherwise. For example, *the mileage of this vehicle is very good*. To eliminate such false collecton we applied a precedence rule: if a collecton talks about a feature (e.g. mileage) along with the overall product (e.g. vehicle) it is considered to describe that feature instead of the product in general.

## 4 Analysis of Sentences

**Handling Conjunctions:** The foremost task needed has been to break the sentences into sub-sentences. This breaking was completely based on the presence of conjunction. Our domain of conjunction consisted of *comma (,)* and *semi-colon (;)* along with *and* and *but*. Qualitative analysis was then performed on the sub-sentences and each sub-sentence contributes to the final rank of the feature. Henceforth the adjectives and adverbs are broken down into sets of *positives* and *negatives*. For the singleton sentences the complete sentence was broken down into sub-sentences and qualitative analysis was performed on each sub-sentence. Each sub-sentence here is looked for the presence of one or more adjective or adverb. In the absence of either of them we presently ignore that sub-sentence.

Once we pick the adjective/adverbs we map them to either positive or negative set and assign the counter as +1 or -1, respectively. We maintained two lists of adjectives corresponding to the sets. The lists have been built by scrutinizing different blogs, and identifying the adjectives/adverbs along with their senses. The actual value of the sub-sentence can be calculated by multiplying all the counters of the qualifiers in that sub-sentence. Table 1 illustrates the above points.

**Table. 1** Analysis of the sentence “The keyboard is very smooth and good looking but the keys are very small.”

Sub-sentence	Adjective	Adverb	Adj Counter	Adv Counter	Net Effect	Remarks
The keyboard is very smooth	Smooth	Very	+1	+1	+1	This qualifies the sentence referring the feature as positive
good looking	Good	-	+1		+1	Another positive comment
the keys are very small	Small	Very	-1	+1	-1	A negative comment

The final cumulative rank for the sentence can simply be calculated by adding the individual rank of each sub-sentence. For example, the rank of the feature “keyboard” in the above example is +1. In case of sub-sentences where the qualifiers do not seem to fall perfectly in either the positive or negative set we

used the value of the previous sub-sentences along with the conjunction they are joined to interpret or validate the present sub-sentence.

Analysis of collecton needs more elaborate schemes. Here it depends both on the type of the conjunction and the feature it is referring to. A particular sub-sentence may refer to a particular feature whereas the other talks about some other feature. Hence in this case one has to identify with each sub-sentence which feature it is referring to. There are two cases of identifying the feature(s) a sub-sentence corresponds to. Firstly, we apply the scheme given in Section 3 for finding out the feature in the case of collecton. This works well if we have at least one feature in each sub-sentence. In the absence of any feature in the sub-sentence we proceed recursively as follows. First we identify the feature present in the initial sub-sentence. For the successive sub-sentences we assume that they refer to the same feature, until we encounter another feature. Once the features are identified the qualitative analysis is performed similar to the singleton sub-sentences.

## 5 Final Rank of the Product

Classifying a product is always a subjective matter concerned with the preferences and liking or need of a particular person. A product may fare well for one user because of the features it provides, but may seem very unappealing to another user. Hence before deciding on the final rank of a product we do need to consider what precisely the user wants out of that product. Depending upon the user's feedback we put weights on the different features of a product. And the final score of the product is calculated as a weighted sum of the scores of the individual features. If a user does not have preferences for certain features their weights are considered to be 0. For the relevant features the weights are normalized so that their sum is 1.

For a particular feature and a particular brand we parse all the relevant blogs and find the total score of a feature. That divided by the number of occurrence of such sentences will give the average quality of the product. We multiply these values with the weight of the feature. This is carried out for each feature to get the final value defining the quality of the product expressed as the requirements provided by the user.

## 6 Experimental Results and Analysis

We present the results for interpretation of laptop and automobile blogs. For each category we have analyzed 20 blogs for the same brand. To validate the result we have compared the software against human interpretation for those set of blogs and reported the percentage of correct conclusion (either positives or negatives

comment) for each feature of the 2 products. Moreover, to assess the efficacy of our algorithm we have also derived the "precision" and "recall" values for each feature and subsequently calculate the F value. These values have been calculated by comparing the output of the algorithm with expert comments.

On an average we have analyzed approximately 1000 sentences in total for each product. We first present the overall correctness of the algorithm. Table 2 shows the percentage of correct responses and percentage of doubtful interpretation given by our software. Here correct means that the human and software both have arrived at the same conclusion (+ve or -ve). The doubtful represents where the human might have arrived at some conclusion but the software is unable to give its judgment. Rest is incorrect interpretation. Due to the limitation of space we show only 4 important features in both the cases. However, the last row of Table 2 and Table 3 shows the average of all the features in the original table.

## 7 Conclusion

In this work we propose an approach for classification of products based on user reviews through interpretation of web-blogs. We find that thorough semantic analysis of the blogs is often difficult, and needs highly efficient NLP techniques. However, by restricting ourselves to specific domains we could retrieve useful sentences, i.e. sentences containing some reviewer's remarks on the product concerned, or some of its features. By considering the adjectives and adverbs used by the reviewer for different features of the product, and noting their senses (positive or negative) we obtain a score for that particular feature of the product. The overall score of the product is calculated as a weighted average of the feature scores, where the weights are determined on the basis of the user's choice.

**Table 2.** Results of laptop blogs

	1 <sup>st</sup> level Interpretation			2 <sup>nd</sup> level Interpretation		
	Correct %	Doubtful %	Wrong %	Precision	Recall	F
IO Devices	65	6	29	0.94	0.72	0.82
Looks	100	0	0	1	0.78	0.875
Multimedia	100	0	0	1	0.74	0.85
Overall	76	24	0	0.875	0.73	0.79
Percentage	84.82	12.90	2.27	97.06	72.78	81.15

**Table 3.** Results of automobile blogs

	1 <sup>st</sup> level Interpretation			2 <sup>nd</sup> level Interpretation		
	Correct %	Doubtful %	Wrong %	Precision	Recall	F
Power	67	0	33	1.00	0.43	0.60
Steering	31	69	0	1.00	0.33	0.50
Engine	57	43	0	1.00	0.58	0.74
Overall	76	9	15	0.76	0.45	0.56
Percentage	65.55	23.45	11	98	45	60

Mining of blogs/texts for review or summarization is not new to NLP practitioners. NLP techniques have been used for movie reviews [6], determining sentiments of opinions [7], and in other domains. Melville et.al. [8] provides a useful discussion on blog analysis and the typical problems arise therein. In this work, we proposed a novel application of blog analysis for ranking products for a prospective buyer.

Our initial experiments reported here show promising results. However, the system is still in its infancy. Proper NLP techniques need to be incorporated into the system for carrying out different tasks, which we have currently done manually. Algorithms need to be developed for identification of key features of a product, automatic identification of senses (which itself is an important NLP task as some terms can be used in both positive and negative sense depending upon the context), etc. Furthermore, the scoring technique needs to be improved. The present strategy of giving +1 and -1 for positive and negative senses respectively appears to be too simplistic. We are currently working in these directions for designing an improved product-ranking system.

## References

1. [www.thetruthaboutcars.com/2008-pontiac-g8-gt-take-two](http://www.thetruthaboutcars.com/2008-pontiac-g8-gt-take-two)
2. [www.carwale.com/Research/ReadFullUserReview-rid-92-sort-1-m-512278949.html](http://www.carwale.com/Research/ReadFullUserReview-rid-92-sort-1-m-512278949.html)
3. James A. : Natural Language Understanding, Pearson, New Delhi (1995)
4. <http://www.laptopmag.com/review/laptops/lenovo-thinkpad-x301.aspx>
5. [www.thetruthaboutcars.com/take-two-2009-honda-accord-lx-review](http://www.thetruthaboutcars.com/take-two-2009-honda-accord-lx-review)
6. Zhuang L., Jing F., Zhu X. : Movie review mining and summarization. In Proceedings of the 15<sup>th</sup> ACM international conference on Information and knowledge management (CIKM), pp. 43–50 (2006)
7. Pang B., Lee L., Vaithyanathan S.: Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2002)
8. Melville P., Gryc W., Lawrence R. : Incorporating Background Knowledge into Text Categorization for Improved Sentiment Analysis. Technical Report, IBM (2008)

# Document Summarization using Wikipedia

Krishnan Ramanathan, Yogesh Sankarasubramaniam,  
Nidhi Mathur and Ajay Gupta  
HP Laboratories  
24, Salarpuria Arena, Hosur Road,  
Adugodi, Bangalore, India  
[{krishnan\\_ramanathan,yogesh,nidhim,ajay.gupta}@hp.com](mailto:{krishnan_ramanathan,yogesh,nidhim,ajay.gupta}@hp.com)

**Abstract.** Although most of the developing world is likely to first access the Internet through mobile phones, mobile devices are constrained by screen space, bandwidth and limited attention span. Single document summarization techniques have the potential to simplify information consumption on mobile phones by presenting only the most relevant information contained in the document. In this paper we present a language independent single-document summarization method. We map document sentences to semantic concepts in Wikipedia and select sentences for the summary based on the frequency of the mapped-to concepts. Our evaluation on English documents using the ROUGE package indicates our summarization method is competitive with the state of the art in single document summarization.

## 1 Introduction

Mobile phones will be the onramp to the Internet for a large fraction of the world's population. However, Internet access on the move often happens in attention deficit situation and the user is capable of assimilating lower amount of information in a mobile context. Hence, the user interaction has to be adapted to the mobile scenario by presenting only the most relevant information to the user. Consequently, many mechanisms have been employed to simplify information presentation on mobile phones of which summarization is one [1]. Most research on document summarization has focused on multi-document summarization; this is more relevant for news sites where documents from multiple news agencies are available. Single document summarization is more relevant to simplifying information consumption on the Internet and on mobile phones, but has received lesser attention [2]. In this paper, we describe a novel, language independent, single document summarization system that uses Wikipedia for sentence selection.

## 2 Related Work

The first paper on summarization appeared in 1958 [7]. Kupiec et al. [9] proposed summarization by sentence extraction in SIGIR 1995. Today, there are broadly four approaches to summarization. The first uses heuristics for rating sentences (e.g. rate sentences that contain document title words higher). The second approach is corpus based and uses TF\*IDF of words in the corpus to identify important words [1]. The third approach uses the structure of text, for instance the method of lexical chains [6]. The final approach is the knowledge-rich approach, our method falls into this category.

Microsoft Word has a summarization algorithm based on a statistical approach. It works only for English documents, this limits its usability. Newsinessence [4] is a multi-document summary of news. A language independent summarization algorithm based on a graph based ranking of sentences (to identify sentence similarity) is presented in [2]. In [1], the authors present five methods for summarizing parts of web pages for handheld devices using a combination of keyword extraction and text summarization.

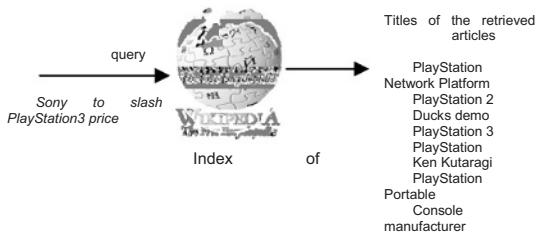
## 3 The Proposed Method

Our approach to summarization is a proxy based approach that processes the document enroute to a mobile device and sends only the summary to the mobile device. Most summarization systems today extract parts of original documents and output them as summaries. Sentence extraction [9] is the most popular way of creating summaries. In this section, we describe a new approach to document sentence extraction using Wikipedia and its application to generating summaries.

### 3.1 Mapping Sentences to Wikipedia Concepts

Wikipedia has grown to become the largest encyclopedia with over 2 million articles.

Our technique is based on using the Wikipedia corpus to find the document topic [10]. We first map individual sentences in the document to Wikipedia concepts. For doing this, the entire Wikipedia corpus is indexed using the Lucene engine. The sentence is then input as a query to the Lucene engine. The titles of the Wikipedia documents are extracted from the results to the query (“hits” in Lucene terminology). This process is illustrated in

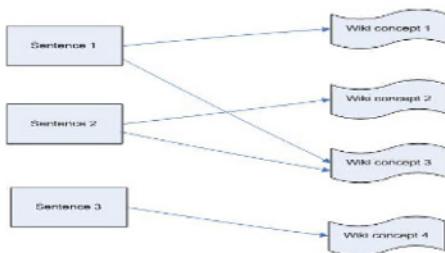


**Fig. 1.** Querying a Wikipedia index

The above step is repeated for each sentence in the document and the number of “hits” for each Wikipedia concept is accumulated in a data structure (C++ multimap).

### 3.2 Construction of the Bipartite Graph

Document sentences are mapped to semantic concepts in Wikipedia by virtue of query “hits” using the Lucene engine as described previously. This mapping can be captured as a bipartite graph, with one set of nodes (or vertices) denoting the document sentences and the other set of nodes denoting the Wikipedia concepts. An edge between a sentence node and a concept node indicates a mapping between the corresponding document sentence and Wikipedia concept, while the absence of an edge indicates that there is no mapping. Figure 2 illustrates this bipartite graph for a small document of three sentences.



**Fig. 2.** Construction of the sentence-Wikipedia concept bipartite graph

Let  $G$  denote the connection matrix of the bipartite graph. The matrix  $G$  is of size  $M \times N$  in general, where  $M$  is the total number of sentences and  $N$  is the number of concepts in the given document. The goal of the summarization algorithm is to use  $G$  to derive the summary  $S$ .

### 3.3 The Summarization Algorithm

After the entire document has been processed and the bipartite graph has been created in the manner outlined above, we identify the Wikipedia concepts that got “hit” multiple times by different sentences in the document. The larger the number of hits, the more that particular concept is relevant to the summary and hence the sentences pointed to by that concept. The sentences in the document that mapped to the Wikipedia concepts with the largest number of hits are selected and output as the summary of the document. In our current system, the user can specify two thresholds for selecting concepts, concepts with hits above the maximum threshold and below the minimum threshold are excluded and sentences that map to them will not be included in the summary.

Figure 2 illustrates the process for a small document of three sentences. Wiki concept 3 is hit by both sentence 1 and 2. Assuming both the min and max threshold was set at 2 hits, sentence 1 and 2 are chosen as the two sentence summary of this three sentence document.

More concretely, we first compute the sum of the columns of the sentence-concept bipartite graph (the in-degree of the columns). We then apply a user specified threshold to select concepts whose column sums are above the specified threshold. These concepts are considered central to the document. This essentially amounts to selecting concepts that are mapped to by the highest number of sentences. For the graph of Figure 2, the matrix is shown below

	C1	C2	C3	C4
S1	1	0	1	0
S2	0	1	1	0
S3	0	0	0	1

The sum of the elements in column 3 is the maximum (equals 2) and hence only concept 3 would be chosen as the representative concept if the threshold was 2 concepts. For column 3, row 1 and row 2 have non-zero entries and hence sentence 1 and sentence 2 are chosen as the summary.

## 4 Evaluation

To give a flavour of the summaries generated by our system, we now reproduce the summary produced by our system for a small document of 13 sentences using the algorithm outlined in section 3.3. This document was chosen from [[http://www.time.com/time/magazine/article/0,9171,985907,00.html?iid=digg\\_share](http://www.time.com/time/magazine/article/0,9171,985907,00.html?iid=digg_share)].

**Original document -**

“Running nose. Raging fever. Aching joints. Splitting headache. Are there any poor souls suffering from the flu this winter who haven’t longed for a pill to make it all go away? Relief may be in sight. Researchers at Gilead Sciences, a pharmaceutical company in Foster City, California, reported last week in the Journal of the American Chemical Society that they have discovered a compound that can stop the influenza virus from spreading in animals. Tests on humans are set for later this year. The new compound takes a novel approach to the familiar flu virus. It targets an enzyme called neuraminidase, that the virus needs in order to scatter copies of itself throughout the body. This enzyme acts like a pair of molecular scissors that slices through the protective mucous linings of the nose and throat. After the virus infects the cells of the respiratory system and begins replicating, neuraminidase cuts the newly formed copies free to invade other cells. By blocking this enzyme, the new compound, dubbed GS 4104, prevents the infection from spreading”.

The summary produced by our system is as follows -

“Are there any poor souls suffering from the flu this winter who haven’t longed for a pill to make it all go away? Relief may be in sight. Researchers at Gilead Sciences, a pharmaceutical company in Foster City, California, reported last week in the Journal of the American Chemical Society that they have discovered a compound that can stop the influenza virus from spreading in animals. The new compound takes a novel approach to the familiar flu virus. It targets an enzyme called neuraminidase, that the virus needs in order to scatter copies of itself throughout the body. After the virus infects the cells of the respiratory system and begins replicating, neuraminidase cuts the newly formed copies free to invade other cells”

Our system produced a summary of five sentences for the original 13 sentence document, all these sentences mapped to just one Wikipedia concept “Influenza”. The sentences marked yellow in the original document are the ones removed by the sentence selection algorithm implemented in the summarizer.

We also evaluated our system on the DUC 2002 single document summarization task. In this task, there are 567 news articles and an expert written summary for each article. The evaluation was done using the ROUGE package [3]. We report the ROUGE\_1 average recall numbers [5] at the 95 % confidence interval and for the first 100 words of each summary; these correspond to the ngram (1, 1) setting of ROUGE. We also used the “-m” option of the ROUGE toolkit for stemming the words. Summaries were generated for different thresholds for the Wikipedia concepts. The results are shown in Table 1 below

**Table 1.** Recall results using ROUGE

Wikipedia concepts threshold	ROUGE recall
2	0.4680
3	0.4586
MAX	0.4368

The Wikipedia concept threshold is the number of hits a concept should receive before it is eligible to vote for a sentence. The MAX evaluation is done by using only a single concept with the highest number of hits. The best system in the DUC 2002 task (system S28) had a ROUGE-1 recall of 0.4804 [11], the above result with 2 concepts would place our system third in the DUC 2002 top performing systems. The recall score dropped with a three concept threshold, this was mainly because 56 documents in this case had zero length summaries (no Wikipedia concept was mapped to 3 times by sentences in these documents).

## 5 Discussion and Future Work

The main limitation of this method is that a sentence could get chosen in the summary by virtue of getting mapped to only one concept. This concept might be a very generic concept pertaining to a high level topic (e.g. Sports). We would like multiple concepts to have a say in whether a sentence should be chosen. The other limitation is that this method does not offer an easy way to control the size of the summary. For instance, if the column score for a concept sums to N and we wish to have a summary of size M where  $M < N$ , the baseline method does not offer a principled way of doing this (we could use heuristics like compute the row sum of the sentences and order them by their row sums). We plan to overcome this limitation by making better use of the bipartite graph. In particular, we plan to devise an algorithm based on the intuition that important sentences in the graph map to important concepts and vice versa. Finally, we wish to evaluate the efficacy of our method in a multi-document summarization scenario such as in [8].

## References

1. Orkut B. et.al.; Seeing the whole in parts: Text summarization for web browsing on handheld devices, WWW 2001 (2001)
2. Mihalcea R., Tarau P. : A language independent algorithm for single and multiple document summarization. IJCNLP (2005)
3. ROUGE package for evaluating summaries, <http://berouge.com/default.aspx>.

4. Newsinessence, <http://lada.si.umich.edu:8080/clair/nie1/nie.cgi>
5. Lin C., Hovy E. :Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of Human Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada (2003)
6. Barzilay R., Elhadad M. : Using lexical chains for text summarization. Proceedings of the ACL workshop on intelligent scalable text summarization, pp.10–17 (1997)
7. Luhn H. : The automatic creation of literature abstract., IBM journal, April (1958)
8. Nguyen P. et.al. : Summarization of multiple user reviews in the restaurant domain. Microsoft Research technical report, MSR-TR-126-2007 (2007)
9. Kupiec J., Pedersen J., Chen F. : A Trainable Document Summarizer. SIGIR (1995)
10. Gabrilovich E., Markovitch S. : Overcoming the brittleness bottleneck with Wikipedia: Enhancing Text Categorization with Encyclopedic Knowledge, Proc. of the AAAI conference (2006)
11. Wan X., Yang J., Xiao J. : Incorporating cross document relationships between sentences for single document summarization, ECDL 2006, LNCS 4172, pp. 403-414, (2006)

# **Improving Performance of English-Hindi CLIR System using Linguistic Tools and Techniques**

Anurag Seetha<sup>1</sup>, Sujoy Das<sup>2</sup> and M.Kumar<sup>3</sup>

<sup>1</sup> Computer Sc. & Applications, MCRPSV, Bhopal, India, anuragseetha@gmail.com

<sup>2</sup> Deptt. of MCA, MANIT, Bhopal, India, sujdas@gmail.com

<sup>3</sup> Deptt. of Computer Sc. & IT, SIRT,Bhopal, India, prof.mkumar@gmail.com

**Abstract.** World Wide Web is growing rapidly and the content on Web of languages other than English is also increasing rapidly compared to English. Hindi is most widely spoken language in India. In past few years Hindi content has also increased rapidly on the Web. To ensure complete information exchange, in the era of globalization the information retrieval systems need to be multilingual or cross lingual. We have designed and developed an English-Hindi Cross Language Information Retrieval (CLIR) System using Dictionary based query translation method. Our previous experiments [5] showed reasonable 64.80% performance of the monolingual retrieval with this system using the TREC style test collection created especially for this research. This paper describes results of the English-Hindi CLIR experiments using some specialized query formulation strategies like stopword removal, stemming of query terms, transliteration of out of vocabulary words etc. The results demonstrated that the performance gradually improved when we applied NLP tools and techniques in short queries. Performance was dropped down to some extent when using query expansion and structuring as well using long queries to obtained cross-language results. The best performance result we obtained from these experiments was 82.91% compared to the monolingual retrieval.

## **1 Introduction**

The goal of information retrieval (IR) is to retrieve documents that are closely related to a user's query. World Wide Web is growing rapidly; further, the content on Web of languages other than English is also increasing rapidly compared to English. In the past few years Hindi content has also increased rapidly on the Web. Almost all major news papers, publication houses and Government departments in the country have setup their web sites in Hindi Language. The globalization is reducing the significance of national borders in terms of trade and information exchange. To ensure complete information exchange, information retrieval systems have got to be multilingual or cross-lingual. Recently, a

considerable research in the information retrieval field is devoted to Cross-Language Information Retrieval. Cross Language (or Cross-lingual) Information Retrieval (CLIR) refers to the retrieval and ranking of documents in one language in response to a query issued in a different language. The retrieval of relevant monolingual information is a difficult task in itself, and the language difference in the CLIR adds an additional dimension to the complexity of information retrieval task. Translating user's query into the document language requires bilingual resources such as machine translation system, bilingual dictionaries, parallel bilingual corpora etc.. "Indian Language Technology Vision 2010" prepared by the Ministry of Information Technology also emphasizes the need of such solutions for Indian languages [1]. We have designed and developed an English-Hindi Cross Language Information Retrieval (CLIR) System. Among several approaches to CLIR, we have used the method of translating user's query into the document language using a bilingual dictionary. When we started this research, a standard Hindi test collection was not available, so we created one for experimentation on the guidelines of TREC, CLEF and NTCIR. We have used publicly available online bilingual dictionary "*Shabdanjali*" from IIIT, Hyderabad [2]. It needed conversion from ISCII [3] to UTF-8 [4] and some normalization to translate the source English query into the target Hindi language query. The dictionary contains approximate 26K entries. The details regarding the design of the system can be found in [5] and the detailed methodology for creating the test collection is given in [6].

## 2 Experimental Setup

To evaluate English-Hindi CLIR, we have setup the experimental environment. We have employed the evaluation methodology being advocated by forums like CLEF, TREC and NTCIR to evaluate a bilingual CLIR system. The details of the experimental settings are given below:

We have used our Hindi document test collection in the experiments. It contains nearly 12000 documents in UTF-8 encoding of which only 6219 documents are used in these experiments with an average length of 399.40 words. A Total of 25 topics are created as per the TREC & CLEF guidelines covers variety of subjects used to evaluate the performance of the system. We have used human accessed binary judgments for the topics obtained from the pooled method.

We have implemented the completely automatic query construction method for the evaluation. Three types of queries are constructed automatically from the topics; the first set composed of the title field alone (T), the second containing the title and description fields (TD), and the third is composed of the title and narrative fields (TN). The type one queries are also called *short queries* and the type two and three are known as *long queries*. The average length of the 25 topics for the Title fields only is 3.6 terms, close to the average topic length used in TREC and Web query average query length [7].

### 3 Preliminary Experiments with the System

Before carrying out cross lingual experiments on the system, we conducted controlled monolingual experiments to set an approximate upper bound on performance. It would not be fair to expect any of automatic query translation to perform better than this approximate upper bound.

We generated and ran automatic queries from these topics on the Hindi document collection to obtain monolingual results. This run entitled the Hindi-Human-Monolingual-run uses queries that were translated in Hindi by language expert. Thus the run can be treated as Hindi monolingual run for all practical purpose. Manual translation of queries and compared its results with the cross-language is now a widely used evaluation strategy. In these run, we used mainly the <Title> field from the topic to generate the queries. The performance statistics of this run for all the queries are shown in the Table 1 under MONO column. These monolingual results give us the upper bound for cross language results.

A query translation needs not be a true translation of the given source language query for IR purpose. In other words, the target language output produced need not be well formed and human readable and the only goal of such a translation is to obtain a topically similar translation in the target language to enable proper retrieval. In our system, a given source language query is translated using word by word translation. However, the design of our system does support phrasal or multi-word expression lookup as well. Each source language word is looked up in the bilingual dictionaries for exact match.

Table 1 describes the various runs we report in this paper. MONO is the Hindi monolingual run while others are English-Hindi cross-language runs using various strategies. The results of experiments with English-Hindi CLIR system using simple word-by-word translation by choosing first, preferred-N, random N and choosing all equivalent from the dictionary have been reported in our previous publication [5] . We have reported that the performance of the system were respectively 64.80%, 57.90% and 57.48%, 11.83% and 57.13% of monolingual retrieval. These results are given in the Table 2 as FIRST, PRE-N-I, PRE-N-II, RND-N and ALLEQV.

The present paper describes more experiments with this system where we have used some specialized query formulation strategies like stop word removal, applying stemming before translation of query, transliteration, query expansion and query structuring etc. They are described in the following sections.

**Table 1.** Run Descriptions

Run ID	Run-Name	Description
Hindi Mono Run	Mono	Using title of the Manually Translated Hindi Topic. This will use as Upper Bound of the system.
Taking First Translation	FIRST	This Cross Language run uses the First translation equivalent from the bilingual dictionary.
Taking Preferred-N translations	PRE-N-I & PRE-N-II	This Cross language run uses the Preferred first N translations from the dictionary for translate the query.
Taking Random Nth Translation	RND-N	In this run random Nth translation equivalent of each is taken for translating the query in each run. If the Nth equivalent is not present then the first equivalent is taken.
Choosing All Equivalent	ALLEQV	This run uses all translation equivalent available in the dictionary for a word for translating the query.
Stopword-Removal	STWR	In this run we remove the stopwords from the query and taking first translation equivalent of remaining words.
Using Stemmer	STEMR	This run uses the stemmer to remove prefixes and suffixes from the given query before translation.
Using Translitrator For OOV	OOV	In this run we use a transliteration scheme to translate the OOV words.
Query Expansion	QRYEXP	In this run we substitutes every term of the title of the topic with their synonyms from a synonym dictionary and then taking
Query Structuring	QUSTRU	In This run structured query is implemented through the use of Boolean operators to address the ambiguity problem.
Long Query - Using Description	DESCR	In this run the Description field of English topic is used to construct the query to determine the effect of query length and the effect of query formulation.
Long Query - Using Narritive	NARR	In this run the Narritive field of English topic is used to construct the query to determine the effect of query length and the effect of query formulation.

#### 4 Effect of Stop Words Removal

Stopwords are those words which are so common and does not serve any useful purpose in query. In this experiment we have removed all the stopwords from the source language i.e. from the English query before translation and taken first translation equivalent from the dictionary for each query term. Those terms of the source language whose equivalents were not found in the dictionary were added to the translated query without translation. To remove the stop-words from the English query a stop-words list for English was used. Results of this Strategy are given in the Table 2 under column STWR. When compared with the

monolingual performance, we obtained the 75.42% performance of the system in cross language.

## 5 Effect of using Morphological Tools

When thinking about possible improvements to an IR system, one of the most obvious areas to be addressed is the morphological variance of words. Users entering the query word ‘work’, will in all likelihood expect a system to retrieve documents containing the words ‘works’ or ‘worked’ as well.

Stemming is a computational process of the NLP field which removing inflectional and derivational affixes and returning a word stem, not necessarily a real word. A *stemmer* is employed to normalise the morphological variants of a word into a common root form. We have implemented porter stemmer in English-Hindi CLIR experiments. We observe that the average precision of the performance increased compared to stopword removal strategy. The results are shown in Table 2 under column STEM.R. This strategy yields 80.76% performance of the systems when compared with the monolingual performance.

## 6 Impact of Out of Vocabulary Terms and Named Entities

One problem seriously affecting CLIR performance is the processing of queries with embedded foreign names, technical terms and proper names, since they are not found in general translation dictionaries, except for the most commonly used terms and names. These out-of-vocabulary (OOV) words are always a problem for dictionary-based CLIR. In typical evaluations, around 50% of out-of-vocabulary words are names. Many proper names and OOV terms remain untranslatable due to limited coverage of translation dictionary. Demner-Fushman and Oard [8] and Al-Onaizan and Knight [9] stated that when using dictionary-based query translation for CLIR, it can be helpful to augment dictionary lookup with some means of processing out-of-vocabulary (OOV) terms.

Early work on dictionary-based approaches to CLIR in European languages generally showed relatively little adverse effect from omission of named entities. One way to solve the above problem is not to rely on a dictionary alone but to adopt automatic translation according to pronunciation similarities, i.e. to map phonemes comprising an English name to sound units of the corresponding Hindi name. This process is called transliteration. In other words transliteration refers to phonetic translation across languages with different orthographies, such as Arabic to English, Japanese to English or English to Hindi [10,11,12]. When the query and document languages have different alphabets, as in our case, English queries and Hindi documents, transliteration may produce a correct Hindi spelling for out-of-vocabulary English names or technical terms.

Since English proper noun and technical words, which were not found in the dictionary do not normally appear unchanged in Hindi text, we explored resolving OOV terms by transliteration, a common practice for person, location, organization names and some technical terms. We developed a transliteration scheme similar to ITRANS [13]. The results using transliterator are shown in the Table 2 under column OOV. We obtained 82.91% performance of our system compared to the monolingual retrieval.

**Table 2.** Run Statistics

METRIC/R UN	Mono	FIRST	PRE-N-I	RND-N	ALLEQV	STWR	STEMR	OOV	QRYEXP	QUSTRU	DESCR	NARR
num_q	25	25	25	25	25	25	25	25	25	25	25	25
num_ret	2415	2086	2109	1546	2109	1886	1992	1992	2100	2015	2093	2400
num_rel	845	820	820	748	820	745	760	760	760	764	844	
num_rel_ret	720	491	468	92	456	497	514	514	252	496	477	469
map	0.5318	0.3446	0.3079	0.0694	0.3038	0.4011	0.4295	0.4409	0.1697	0.4033	0.3633	0.3243
R-prec	0.5329	0.3452	0.3030	0.0778	0.2849	0.3897	0.4167	0.4263	0.1717	0.3880	0.3631	0.3376
bpref	0.8434	0.5271	0.4753	0.1118	0.4446	0.5847	0.6411	0.6498	0.4365	0.5996	0.5758	0.5149
recip_rank	0.6885	0.5271	0.4782	0.2432	0.5130	0.6180	0.6177	0.6362	0.3199	0.5741	0.5589	0.5500
ircl_prn.0.00	0.7250	0.5408	0.5049	0.2438	0.5148	0.6217	0.6212	0.6397	0.3766	0.6041	0.5613	0.5785
ircl_prn.0.10	0.6867	0.4761	0.4370	0.1134	0.4369	0.5535	0.5694	0.5879	0.3255	0.5529	0.5176	0.5216
ircl_prn.0.20	0.6576	0.4529	0.4254	0.1076	0.4238	0.5294	0.5530	0.5715	0.2705	0.5497	0.5005	0.5057
ircl_prn.0.30	0.6242	0.4488	0.4171	0.1070	0.4148	0.5208	0.5444	0.5563	0.2389	0.5255	0.4825	0.4770
ircl_prn.0.40	0.6140	0.4080	0.3927	0.1065	0.3693	0.4720	0.5011	0.5122	0.2166	0.4974	0.4280	0.4295
ircl_prn.0.50	0.5855	0.3923	0.3759	0.0978	0.3528	0.4546	0.4845	0.4961	0.1861	0.4816	0.3871	0.3854
ircl_prn.0.60	0.5486	0.3528	0.3324	0.0673	0.3135	0.3966	0.4297	0.4413	0.1483	0.4231	0.3498	0.3057
ircl_prn.0.70	0.5322	0.3201	0.2905	0.0000	0.2732	0.3541	0.3885	0.4017	0.1436	0.3714	0.2973	0.2414
ircl_prn.0.80	0.4484	0.2628	0.2159	0.0000	0.2042	0.2966	0.3332	0.3468	0.1096	0.2998	0.2606	0.0919
ircl_prn.0.90	0.3501	0.1618	0.0772	0.0000	0.0709	0.1818	0.2208	0.2367	0.0695	0.1462	0.1798	0.0745
ircl_prn.1.00	0.1155	0.0111	0.0091	0.0000	0.0091	0.0122	0.0593	0.0751	0.0275	0.0730	0.0702	0.0221
P5	0.5840	0.4273	0.3636	0.1556	0.4182	0.5100	0.5238	0.5333	0.2190	0.4571	0.4545	0.4583
P10	0.5680	0.4318	0.3727	0.1444	0.3955	0.5150	0.5333	0.5381	0.2143	0.4714	0.4545	0.4375
P15	0.5547	0.4242	0.3667	0.1296	0.3879	0.4933	0.5143	0.5270	0.2222	0.4730	0.4576	0.4361
P20	0.5420	0.4045	0.3523	0.1194	0.3727	0.4775	0.4881	0.4976	0.2190	0.4500	0.4273	0.4125
P30	0.5200	0.3758	0.3333	0.1241	0.3303	0.4400	0.4429	0.4476	0.1873	0.4159	0.3848	0.3653
P100	0.2880	0.2232	0.2127	0.0511	0.2073	0.2485	0.2448	0.2448	0.1200	0.2362	0.2168	0.1954
P200	0.1440	0.1116	0.1064	0.0256	0.1036	0.1242	0.1224	0.1224	0.0600	0.1181	0.1084	0.0977
P500	0.0576	0.0446	0.0425	0.0102	0.0415	0.0497	0.0490	0.0490	0.0240	0.0472	0.0434	0.0391
P1000	0.0288	0.0223	0.0213	0.0051	0.0207	0.0248	0.0245	0.0245	0.0120	0.0236	0.0217	0.0195

## 7 Query Expansion

The problem of word mismatch is fundamental to information retrieval. Simply stated, it means that people often use different words to describe concepts in their queries than authors use to describe the same concepts in their documents. An obvious approach to solve this problem is query expansion. The query is expanded using words or phrases with similar meaning to those in the query and the chances of matching words in relevant documents are therefore increased. Query expansion has been shown to improve CLIR effectiveness in [14, 15].

The query expansion techniques are of two types:

1. pre-translation query expansion, and
2. post-translation query expansion

The query expansion, we used, substitutes every term of the title of the topic with their synonyms from a synonym dictionary. Synonyms play an important role in concept based query formulation and expansion [16], and in replying to free form queries [17].

The results are shown in Table 2 under QRYEXP column. When compared with the monolingual performance, we obtained the 31.91% performance of the system in cross language. The results show that the pre-translation query expansion in synonyms are added to the source language query does not add useful terms to the queries and therefore the performance is lower than the base results.

## 8 Query Structuring for Disambiguation

Query-by-query analysis of resultants translated queries in the experiments showed the ambiguity in the meaning. Some previous researches have also demonstrated that ambiguity is an inherent problem with the dictionary-based query translation. Bilingual transfer dictionaries are an important resource for query translation in cross-language text retrieval. However, term translation is not an isomorphic process, so dictionary-based system must address the problem of ambiguity in language translation.

We have implemented structured query technique to address the ambiguity and the translation problem by taking advantage of the structure in a Boolean query. The Boolean model provides a natural method for reducing the impact of spurious translations by enforcing conjunctive relationships between query terms. Boolean disjunction (the OR operator) is a natural way to link together many translation equivalents without dramatically increasing the weight of the underlying concept. Therefore, our CLIR system connects translation equivalents of the same term with the OR operator. Boolean conjunction (the AND operator) is likely to be an effective strategy for disambiguation if one believes the following hypothesis: "The correct translation equivalents of two or more query terms are much more likely to occur together in target language documents than in conjunction with any incorrect translation equivalents." One can think of incorrect translations as (relatively) random noise added to the query. The chance that medium to low frequency noise terms (derived from different source language words) will occur together regularly is remote queries.

The summary statistics for all the topics of this strategy are shown in the Table 2 under QUSTRU column. When compared with the monolingual performance, we obtained the 75.84% performance of the system in cross language.

## 9 Query Length and its Effect on Performance

In all the above experiments, we used only the ‘title’ fields of the topic for query generation, which contains only few words. Often users enter long natural language sentences to express his/her information need. In our next set of experiments, we used that parts of the topics where the requirements are expressed in sentences. It would be beneficial to determine the effect of query length and the effect of query formulation using our current approach. Here queries are constructed using fully automatic query construction method that uses the <description> and <narrative> fields. In addition to stop word removal, repeated phrases like “Document should give information related to”, “A relevant document must report on” etc. are also removed from the description and narrative.

### 9.1 Using Description Field

In this run the description field of English topic is used to construct the query. Results shown under column DESCR of Table-I shows the summary statistics for all the topics of this strategy. It may be noted that when compared with the monolingual performance, we obtained the 68.32% performance of the system in cross language.

### 9.2 Using Narrative Field

In this run the narrative field of English topic is used to construct the query. Results shown under column NARR of Table-I shows the summary statistics for all the topics of this strategy. When compared with the monolingual performance, we obtained the 60.98% performance of the system in cross language.

## 10 Conclusion

English-Hindi CLIR is an interesting area of research in Human Language Technology and has great utility. The results we have obtained using NLP tools and resources in all show significant improvement over the previous runs. This research demonstrates query translation methodology with simple resources and some NLP tools and integrates it with the existing IR engine to provide a workable solution to CLIR. Present experiments were carried to study the performance evaluation of the English-Hindi CLIR system under some specialized query translation and construction strategies. Following observations have been made:

- When we removed stop-words from the source language query in the run STWR the performance became 75.42% of the monolingual performance,

compared to 64.80% when first translation equivalent of the dictionary was taken to compare with the monolingual performance.

- When we removed stop-words from the source language query and also applied the stemming process in the run STEMR, the performance became 80.76% of the monolingual performance, compared to 75.42% when first translation equivalent of the dictionary was taken. We have also observed that some of the query terms are sometimes wrongly over-stemmed or under-stemmed during the suffix removal process which might have led to wrong translation results.
- When we used the transliteration OOV words in the run OOV, the performance became 82.91% of the monolingual performance, compared to 80.76% of the run STEMR.
- When we tried to expand the query with the source language synonyms word from a separate synonym dictionary, the performance drastically dropped to the level of 31.91% of the monolingual query in the run QRYEXP. This is because the query expansion process from synonym adds many irrelevant concept words to the source level query, thus this strategy is not useful in the enhancement of the retrieval performance.
- The structuring of the query using Boolean operators (run QUSTRU) gives the performance 75.84% of the monolingual, in comparison to the 82.91%, achieved in the run OOV.
- The run DESCRI, which uses the description and the narrative (run NARR) fields of the topic to generate the automatic queries for retrieval gives the performance 68.32% and 60.98% respectively of the monolingual average precision. This is very less in compared with the strategy OOV, which gives the best results.

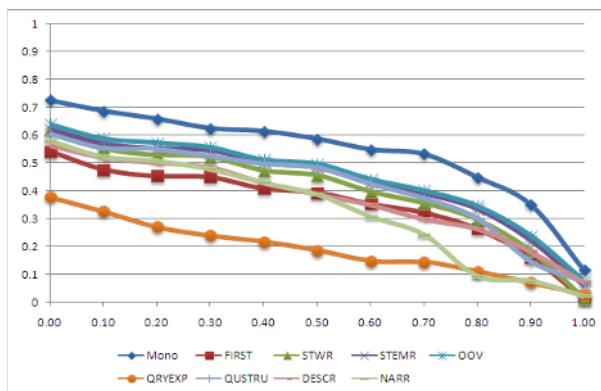
It may be noted that the run OOV that used the first translation equivalent from the dictionary with stop word removal, stemming and transliteration of OOV gave the best results in English-Hindi CLIR.

In order to analyse the overall effectiveness of our query translation approach for all 25 queries, a comparison of all the retrieval strategies used in the evaluation of the system is given in Table 3.

**Table 3.** Comparative Performances of the English-Hindi CLIR system in all strategies (Based on Average Precision)

Strategy	Average Precision	% of Mono
MONO	0.5318	--
FIRST	0.3446	64.80
PRE-N-I	0.3079	57.90
PRE-N-II	0.3057	57.48
RND-N	0.0629	13.05
ALLEQV	0.3038	57.13
STWR	0.4011	75.42
STEMR	0.4295	80.76
OOV	0.4409	82.91
QRYEXP	0.1697	31.91
QUSTRU	0.4033	75.84
DESCR	0.3633	68.32
NARR	0.3243	60.98

Comparison with the monolingual upper bound is the preferred measure employed in the field and our results compare well with the best reported results in the literature. However, our manual analysis and investigation of individual query result suggests that there are gaps in the performance our CLIR system over different individual query topics. Though our results on an average show nearly 82% success in comparison with the monolingual performance, there is still scope for improvement. The Figure 1 shows the comparative Recall Precision of all strategies.



**Fig. 1.** Comparative Recall-Precision graph of monolingual and various cross-language runs

## References

1. Ministry of Information Technology, New Delhi, India.: TDIL Vision 2010. Vishwabharat@TDIL Newsletter of Technology Development for Indian Languages (TDIL). January, 2003. (2003)
2. <http://ltrc.iiit.net/>
3. The ISCII document IS13194:1991, Bureau of Indian Standards, BIS (1991)
4. The Unicode Standard, Version 4.0, <http://www.unicode.org>
5. Seetha, A., Das, S., Kumar, M.: Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method. In : Proceedings of 10th International Conference on Information Technology (ICIT 2007), <http://doi.ieeecomputersociety.org/10.1109/ICIT.2007.40> (2007)
6. Seetha, A., Das, S., Kumar, M. : Construction of Hindi test collection for CLIR research. In: Proceedings of International Conference on Cognitive Systems (ICCS 2004). New Delhi, December 14-15 (2004)
7. Jansen, B., Spink, A., Saracevic, T.: Real life, real users, and real needs: A study and analysis of user queries on the Web. *Information Processing and Management*. 36(2), pp 207–227 (2000)
8. Demner-Fushman, D., Oard, D. W.: The effect of bilingual term list size on dictionarybased cross-language information retrieval. In: 36th Annual Hawaii International Conference on System Sciences (HICSS'03) - Track 4. Hawaii. (2003)
9. Al-Onaizan, Y., Knight, K.: Machine Transliteration of Names in Arabic Text. In: Proceedings of ACL workshop on Computational Approaches to Semitic Languages.
10. Knight, K., Graehl, J.: Machine Transliteration. *Computational Linguistics*. 24 (4), 599–612. (1998)
11. Stalls, B. G., Knight, K.: Translating Names and Technical Terms in Arabic Text. In: Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages. pp. 34–41, Montreal: ACL. (1998)
12. Qu, Y., Grefenstette, G., Evans, D. A.: Automatic Transliteration for Japanese-to-English Text Retrieval. In: Proceedings of the 26th Annual Inter-national ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 353–360. New York: ACM Press. (2003)
13. ITRANS Indian language transliteration package at [www.aczone.com/itans](http://www.aczone.com/itans).
14. Adriani, M., van Rijsbergen, C. J. : Term Similarity Based Query Expansion for Cross Language Information Retrieval. In: Proceedings of Research and Advanced Technology for Digital Libraries. Third European Conference (ECDL'99). pp. 311–322. Springer Verlag: Paris, September (1999)
15. Ballesteros, L., Croft, W. B.: Resolving Ambiguity for Cross-language Retrieval. In: Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval. (1998)
16. Kekäläinen, J., Järvelin, K.: The impact of query structure and query expansion on retrieval performance. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, (1998)
17. Kristensen J.: Expanding end-users query statements for free text searching with a search-aid thesaurus. *Information Processing and Management* 29(6), 733–744

# **Improving Multi-document Text Summarization Performance using Local and Global Trimming**

Kamal Sarkar

Computer Science & Engineering Department,  
Jadavpur University, Kolkata – 700 032, INDIA  
jukamal2001@yahoo.com

**Abstract.** Multi-document summarization can produce a condensed representation of the contents of multiple related text documents. With this summarization facility, web users can judge rapidly the relevance of a group of documents returned by the search engines and decide whether those should be discarded. This reduces the total search cost for the users. This paper presents a multi-document summarization system, which has two components: (1) the sentence extraction component that produces draft summaries by sentence extraction and (2) the sentence-trimming component that eliminates the low content and redundant elements from the sentences in the draft summaries for improving the summarization performance. In this paper, we also introduced several new local and global sentence-trimming rules. Our experiment on DUC 2004 data set shows that the local and global trimming can improve the extractive multi-document summarization performance in many cases.

## **1 Introduction**

The web users are overwhelmed with a large volume of information even on a single topic returned by the traditional search engines and it is very difficult for the users to go through all the hits and find the relevant information from the collection. Multi-document summarization is a process, which produces a condensed representation of the contents of multiple related text documents collected from heterogeneous sources for human consumption. So, Multi-document summarization facilitates human to digest the main contents of multiple related text documents very rapidly. With this kind of summarization facility, users can discard a set of documents after going only through the summary (gist) of them if they are not relevant to them. Thus the total search cost is reduced.

Previous work on extractive summarization ranks sentences based on simple features such as their position in the text, term frequency (TF), or some key phrases indicating the importance of the sentences [1, 2, 3] and select top n sentences based on the compression ratio. In another approach to multi-document summarization, information extraction was used to identify similarities and differences across the documents in the set [4].

Redundancy is one of the important factors in multidocument summarization. Some systems ranks sentences based on some sentence-level and word-level features and selects the top most sentence first and measure the similarity of a next candidate textual unit (sentence or paragraph) to that of previously selected ones and retain it only if it contains enough new (dissimilar) information. A popular such measure is maximal marginal relevance [5].

The clustering-based approach controls redundancies in the final summary by clustering sentences to identify themes of common information and selecting one representative sentence from each cluster in to the final summary [6, 7].

Centroid-based multi-document summarization [8, 9] ranks sentences based on its similarities to the centroid which is a pseudo-document consisting of words with TF\*IDF scores greater than a predefined threshold. Here, TF means term frequency, which has been computed by the average number of occurrences of a word across the set of the documents to be summarized and IDF (used in information retrieval task) means inverse document frequency signifies rarity of a word in a text corpus.

Few approaches use information fusion techniques to identify repetitive phrases from the clusters and the phrases are fused together to form the fluent summary [10].

The majority of the summarization systems are either (1) sentence extraction based or (2) using sentence extraction as the primary component of the system. Some researches have been initiated to address the possibility of improving document summarization performance through summary revisions [11]. A pilot study on improving the sentence extraction based summarization performance by sentence compression has been presented by Lin in [12]. In this study, it has been reported that local optimization at sentence level even using a good compression algorithm proposed by Knight and Marcu [13] is not enough to boost the system performance because the basic goal is to find the best compressed summaries not the best compressed sentences. They have suggested that global cross sentence optimization may boost the system performance, but they did not present any global optimization method in this paper. A BE (basic element) -based sentence compression algorithm has been presented in [14].

Compared to the previous work, our multi-document summarization system can be described as extractive summarizer with some local and global trimming capabilities. Sentence trimming is the main focus of our work. The sentence trimming component of this summarization system accepts the draft summaries as input and eliminates the inessential, low content and redundant elements from the sentences in the draft summaries using simple local and global syntactic sentence trimming rules for making room for more number of informative words to appear in the fixed size final summary. The draft summaries are tagged by a POS tagger [15] before its submission to the trimming component. The sentence-trimming component removes phrase level redundancies using trimming rules. Some previous local sentence compression methods [16] use the parsed sentences as input to a trimmer whereas we use POS tagged sentences as input to the proposed trimmer. Reason behind choosing tagger in place of a parser is that the tagger is relatively faster than parser and the taggers that use the same tagsets show the relatively less variations in its outputs. In comparison to the BE (basic elements)-

based sentence compression, our sentence compression technique is simple and fast because the BE-based sentence compression algorithm requires breaking sentences in to a number of basic elements using BE package [17] and parsing the sentences. Moreover, BE-based sentence compression algorithm concentrates only on removing the redundant BEs (basic elements) from the summary.

In the section 2 we describe the a variant of centroid based sentence extraction method. Local and global sentence trimming rules have been covered in section 3. The experiments and results are discussed in section 4.

## 2 Sentence Extraction

We use a variant of a centroid based sentence extraction method [8] for sentence extraction. Our sentence extraction component has a number of steps: (1) Document preprocessing (2) Centroid based sentence ranking (3) draft summary generation.

**Preprocessing.** The preprocessing task primarily includes stop word removal, handling abbreviations, which may contain dots (.). The dots in abbreviations and numeric words (ex. 12.5 millions) may mistakenly be recognized as a sentence boundary. We have used a number of syntactic rules to differentiate between dots in abbreviations, numeric words and dots at the end of the sentence. We replace the dot by the special character (^). We also remove commas (,) from the numeric words (ex. 12,000 people). Before applying sentence-ranking algorithm, input text files contain dots only at the end of the sentence boundary. So, input documents are now easily broken in to a cluster of sentences by considering dots as the sentence separators.

In the final summary, we replace the special characters (^) by dots to transform the words in to original form.

**Centroid Based Sentence Ranking.** The sentences are ranked based on their similarities to the centroid. Here, we consider that a centroid is a pseudo-document consisting of words with  $\log(1+TF)$  scores greater than a predefined threshold and TF (term frequency) = the total number of times a word occurs in the input collection of the documents to be summarized. Like centroid-based summarization, we consider that the sentences containing more words from the centroid of the cluster are more central. This is a measure of the closeness of the sentence to the centroid of the cluster. The score of a sentence S is computed as follows:

$$\text{Score}(S) = \sum \text{Weight}(w_i), \text{ where } \text{Weight}(w_i) = \log(1+TF)$$

$$w_i \in S$$

$w_i$  is a word which belongs to the sentence  $S$  and whose weight is above a predefined threshold.

**Draft Summary Generation.** After ranking the sentences based on their scores, the top ranked sentence is selected first and continues choosing sentences until the desired summary length is reached. While a sentence is selected it is compared to the already selected sentences to verify whether this sentence contain sufficient dissimilar information. To do so, we used a similarity measure between the sentence under consideration and the already selected sentences. If in any case it is found that the similarity value is greater than a threshold value, the sentence is not included in to the summary. The similarity measure used in this work is as follows:

$$\text{Sim}(S_i, S) = (2 * |S_i \cap S|) / (|S_i| + |S|)$$

where  $S_i$  is a sentence belongs to the set of already selected sentences and the  $S$  is a sentence under consideration.

We generate a draft summary of 200 words with the target of generating the final summary of 665 bytes (approximately 100 words).

### 3 Sentence Trimming

A draft summaries of 200 words are generated by the sentence extraction component. Then, the draft summaries are tagged by a POS tagger. The initial summary of 200 words is divided into two parts: size of the first part is 665 bytes (target size of the final summary) and the rest is considered as the second part. If the draft summary of 665 bytes includes a fragment of a sentence, we include the sentence as a whole in the draft summary. The sentences in the first part of the draft summary are reordered using time order and text order [18]. The trimming rules are applied on the sentences in the first part of the draft summary and spaces freed by the trimming are filled with the words selected sequentially from the top of the second part of the draft summary. Thus, a final summary of 665 bytes is generated. We reorder the trimmed summary again to increase readability.

For trimming, the first part of the tagged draft summaries are scanned from top to bottom and the trimming rules are tried on one sentence at a time. We categorize the trimming rules as local and global trimming rules. When we remove low content words from a sentence, we call it as local trimming and when we remove a redundant constituent from a sentence in consultation with other sentences in the draft summary we call it as global trimming.

### 3.1 Local Trimming

The local trimming rules and their application to the English sentences are shown below (Rules are written in italics font and numbered as R1, R2 etc.)

*R1: Delete time words and time expressions.*

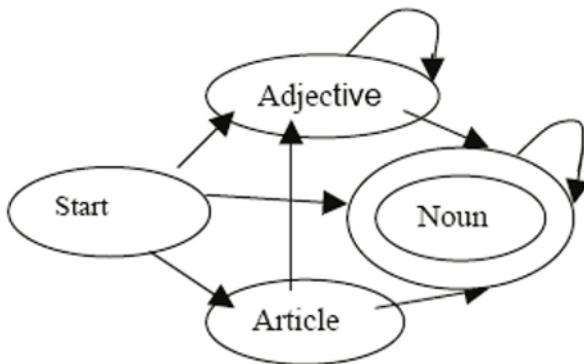
The following are some examples of the time-word sequences that have been observed by us. "On Sunday", "Sunday", "Sunday night", "By late Sunday", "since last Tuesday", "last year", "last week", "last month", "the first time", "next month", "this year"

*R2: Remove adverb directly from the sentences.*

The adverb is tagged by the tagger by the tag <RB>. The constituents to be deleted in the following sentence are shown in bold italic.

[Portuguese/JJ writer/NN who/WP took/VBD up/IN literature/NN **relatively/RB** late/JJ in/IN life/NN and/CC whose/WP\$ **richly/RB** imaginative/JJ . .]

*R3: Remove adjectives with idf <2.5 from the noun phrase if the noun phrase contains more than 2 keywords. Here idf means inverse document frequency (used in traditional information settings). The DFA used for identifying noun phrases is shown in Fig-1.*



**Fig.1.** DFA for noun phrase identification

The constituents, which can be deleted in the following tagged-sentence using rule R3 is shown in bold italic.

[But/CC with/IN exorbitant/JJ salaries/NNS paid/VBN to/TO **several/JJ** unproven/JJ stars/NNS over/IN the/DT **last/JJ** few/JJ years/NNS...]

*R4: Delete the news source information (such as a "Thai newspaper reported Sunday") mentioned at the end of the sentence.*

This segment usually starts with a comma (,) and ends with a period(.). This segment is identified by a list of domain specific keywords and phrases such as "reported", "announced", "according to", "officials said".

**R5:** *If the sentence starts with a PP, delete it if noun heads in the PP are lightweight.*

Here, lightweight means the weight is less than a threshold (2.2 in this setting). The constituents, which can be deleted in the following tagged sentence using rule R5 is shown in bold italic.

[**Along/IN the/D<sub>T</sub> way/NN in/IN an/D<sub>T</sub> almost/RB casual/JJ fashion/NN,/**, the/D<sub>T</sub> document/NN seems/VBZ to/TO confirm/VB two/CD of/IN the/D<sub>T</sub> central/JJ...]

### 3.2 Global Trimming

Modifiers are used before or after the named entities such as person name, organization name and location name. The noun phrases also contain modifier terms such as adjectives, adverbs. The modifier terms may repeat partly or with its entirety in the summary along with the units (NE or NP) it is associated with. The redundant information found in such cases can be deleted with minimum loss of information and without loss of grammaticality. We consider the modifier terms as candidates for global trimming and the parts of modifiers can be eliminated if it is already found in the previously selected sentences. We apply different syntactic rules to identify modifier terms from the noun phrases and from the surroundings of the named entities. So, we have two types of global trimming: Named Entity centric sentence trimming and sentence trimming by thinning NPs.

**Named Entity Centric Trimming.** Named Entity centric trimming has three steps: named entity identification and identification of modifiers surrounding named entities and formation of trimming rules

A word with tag <NNP> (which is used by the tagger to indicate proper noun) has been considered as a part of a named entity and a sequence of words tagged with <NNP> constitutes a named entity.

In our work, we consider a noun phrase having named entity (NE) at the head as a named entity phrase (NEP). An example of NEP is “former Chilean dictator Augusto Pinochet”, where “Augusto Pinochet” is a named entity which is at the head of NEP.

We divide NEP into two parts as < modifier+ NE>. But, sometimes it may happen that a very common word (such as "President") appears before a named entity and tagger uses the tag <NNP> to label it since the first letter of the word is capital. To handle this situation, we check the idf value of this word. If the word is found in the vocabulary and idf value is <2.5, we consider this word as a part of the modifier, otherwise we consider it as a part of the named entity. The procedure to identify a modifier and named entity works as follows:

Say, A is the left most word and H is a head - word in a NEP. Scan NEP from right to left checking NNPs and stop exactly when we encounter a non-NNP or a NNP with idf value<2.5. Say, the word is X. Then we consider the segment spanning the word A to word X as modifier and the rest of the NEP is considered

as a NE. A modifier of a named entity may appear as “noun in apposition” and this form of modifier are extracted using syntactical patterns <NE, M, > or <NE, M.> where M is the modifier and NE is the named entity. Rule for named entity centric phrase trimming is as follows:

*R6A: Delete the modifier of a named entity in its current mention in a sentence if the modifier of its current mention is similar to one of modifiers of its early mentions in already-scanned sentences.*

To apply the above rule, we maintain a list of modifiers for each mention of a named entity covered in the previously selected sentences while scanning the draft summary from top to bottom. Two modifiers are taken to be similar when term based similarity between them is greater than a threshold value (which is 0.5 in our setting). Similarity between two modifiers  $M_1$  and  $M_2$  is calculated as follows:

$$\text{Sim } (M_1, M_2) = (2 * |M_1 \cap M_2|) / (|M_1| + |M_2|)$$

Articles and prepositions (if any) are removed while measuring similarity between two modifiers. Candidate segment for trimming using rule R6A is shown below in bold italics and the similar phrase in the previously scanned sentence is shown in bold only.

Tagged Input:

/Saudi>NNP exile/NN Osama>NNP bin/NN Laden>NNP ,/ the/DT alleged/VBN mastermind/NN of/IN a/DT conspiracy/NN to/TO attack/VB . . . ]  
 [. . . interview/NN of/IN **Afghanistan>NNP** -/: **based/VBN** **Saudi>NNP** **billionaire>NN** **Osama>NNP** **Bin>NNP** **Laden>NNP** who/WP has/VBZ been/VBN accused/VBN...]

After application of rule R6A:

/Saudi>NNP exile/NN Osama>NNP bin/NN Laden>NNP ,/ the/DT alleged/VBN mastermind/NN of/IN a/DT conspiracy/NN to/TO attack/VB . . . ]  
 [. . . reproducing/VBG a/DT foreign/JJ newspaper/NN interview/NN of/IN **Laden>NNP** who/WP has/VBZ been/VBN accused/VB . . . ]

According to similarity metric mentioned above, similarity between two modifiers of NE head “Laden” is  $>0.5$ . So, the modifier of the entity “Laden” in the second sentence has been deleted according to the rule R6A.

*R6B: If any NP matches completely (word by word) with the modifier of a NEP already seen in the previously scanned sentences, replace the NP with NE head of the NEP.*

Tagged input:

[. . . the/DT trial/NN of/IN Malaysia>NNP former/JJ deputy/NN prime/JJ minister/NN Anwar>NNP Ibrahim>NNP on/IN charges/NNS of/IN corruption/NN . . . ]

[. . . because/IN of/IN his/PRP\$ concerns/NNS about/IN the/DT arrest/NN of/IN ***Malaysia/NNP s/PRP former/JJ deputy/NN prime/JJ minister/NN ./]***

After application of Rule6B:

[. . . the/DT trial/NN of/IN ***Malaysia/NNP former/JJ deputy/NN prime/JJ minister/NN Anwar/NN Ibrahim/NNP*** on/IN charges/NNS of/IN corruption/NN . . .]

[. . . , because/IN of/IN his/PRP\$ concerns/NNS about/IN the/DT arrest/NN of/IN ***Anwar/NNP Ibrahim/NNP***]

**Simple NP Trimming.** We trim the noun phrases containing a named entity at its head using Named Entity centric trimming rules discussed above. But, We treat differently with the noun phrases having no named entity at its head. In this case, we consider the trailing non-noun words (adjectives and adverbs) of a noun phrase as modifier terms if the length of the NP is  $>2$  and the distance between the word and the noun head is  $\geq 2$ . The distance between two phrasal words is measured by position of the head-word in the phrase minus position of a word in the phrase. Position value of a word in a phrase increases from left to right of the phrase. The noun words satisfying above syntactical constraints is considered as a modifier words if idf value of the word is  $<2.5$ . The low idf value of the word signifies that the word is very common in the text corpus. The noun phrases of this kind are trimmed using the following rule.

*R7: If A is a NP in a sentence S and B<sub>i</sub> is one of NPs belonging to the list of noun phrases found in the already-scanned sentences and head (A)=head (B<sub>i</sub>), delete the modifier words of A, which matches with that of B<sub>i</sub>*

Candidate segment for trimming using rule R7 is shown below in bold italics and the similar phrase in the previously scanned sentence is shown in bold only. The words to be trimmed are underlined.

[. . .the/DT first/JJ component/NN of/IN ***a/DT multibillion/JJ dollar/NN international/JJ space/NN station/NN*** after/IN a/DT year/NN of/IN delay/NN]  
 [The/DT first/JJ part/NN of/IN ***the/DT international/JJ space/NN station/NN*** was/VBD smoothly/RB orbiting/VBG Earth>NNP. . .]

## 4 Experiments and Results

We chose as our input data the document sets used in the task2 for the evaluation of multi-document summarization during the Document Understanding Conference (DUC) in 2004. This collection contains 50 test document sets, each with approximately 10 news stories. For each document set four human-generated summaries are provided for the target length of 665 bytes (approximately 100 words).

We adopted an automatic summary evaluation metric for comparing system-generated summaries to reference summaries written by humans. The method,

ROUGE [19], is based on n-gram overlap between the system-produced and reference summaries. As such, it is a recall-based measure, and it requires that the summary length be controlled to allow meaningful comparisons. ROUGE reports separate scores for 1, 2, 3, and 4-gram matching between the model summaries and the generated summary. Among these different scores, unigram-based ROUGE score (ROUGE-1) agrees most with human judgements.

In our experiment, for each input data set, a draft summary of 200 words is generated by the sentence extraction component. Then, the draft summaries are tagged by a POS tagger. The initial summary of 200 words is divided into two parts: size of the first part is 665 bytes (target size of the final summary) and the rest is considered as the second part. The trimming rules are applied on the sentences in the first part of the draft summary and space freed by the trimming is filled with the words selected sequentially from the top of the second part of the draft summary. Thus, after trimming and resizing, a final summary of 665 bytes is generated. The results of the evaluation of the overall summarization performances using ROUGE package is shown in table 1. The summarization performances before trimming and after trimming and resizing are shown separately in the table in terms of ROUGE-1 scores. It also shows that trimming improves the summarization performance.

**Table 1.** ROUGE –1 scores for summaries on DUC2004 data with or without applying trimming

	Min	Max	Average
Sentence extraction (without trimming)	0.3536	0.3856	0.3703
Sentence extraction + trimming	0.3601	0.3926	0.3765

The ROUGE [20] package that we used, gives minimum, maximum and average ROUGE scores of all the summaries submitted for evaluation. To evaluate the effectiveness of trimming, we need to evaluate each individual generated-summary before trimming and after trimming and resizing. For this summary evaluation, we compute ROUGE-N scores using the formula given in [19]. ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries, where  $n$  stands for the length of n-gram ( $n=1$  for unigram,  $n=2$  for bigram so on). Out of 50 cases, trimming and resizing of the draft summaries improves summarization performances in 30 cases (60%), the performance remains the same in 10 cases (20%) and the performance slightly degrades in 10 cases (20%).

## 5 Conclusion

In this paper, we have shown that sentence extraction based summarization performance can be improved using local and global trimming rules. We have used only syntactic trimming rules to eliminate less important or redundant constituents of the summary sentences. The more improvement in the overall summarization performance is possible by introducing new trimming rules and its successful application to compress the draft summary. However, these parts will be investigated in future.

## References

1. Baxendale, P. B.: Man-made index for technical literature—An experiment. IBM Journal of Research and Development 2(4), 354–361 (1958)
2. Edmundson, H. P.: New methods in automatic extracting. Journal of the Association for Computing Machinery 16(2), 264–285 (1969)
3. Luhn, H. P.: The automatic creation of literature abstracts. IBM Journal of Research Development 2(2) , 159–165 (1958)
4. McKeown, K. R. and Radev R.D.: Generating summaries of multiple news articles. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Seattle, July, pp. 74–82 (1995)
5. Carbonell, Jaime G. and Goldstein, J.: The use of MMR, diversity-based re-ranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: Melbourne, Australia, pp.335–336 (1998)
6. McKeown, K., Klavans J., Hatzivassiloglou V., Barzilay R., and Eskin, E.: Towards multi-document summarization by reformulation: Progress andprospects. In Proceedings of the 16th National Conference of the American Association for Artificial Intelligence:, pp. 453–460, 18–22 July (1999)
7. Marcu, D and Gerber L.: An inquiry into the nature of multi-document abstracts, extracts, and their evaluation. In Proceedings of the NAACL-2001 Workshop on Automatic Summarization: Pittsburgh, June. NAACL, pages 1–8 (2001)
8. Radev, D. R., Jing, H., Budzikowska. M. Centroid-based summarization of multiple documents: Sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACL Workshop on Summarization: Seattle, April (2000)
9. Radev, D. R., Jing, H., Sty M., Tam, D.: Centroid-based summarization of multiple documents. Inf. Process. Manage. 40(6), 919-938 (2004)
10. Barzilay, R., McKeown, K., Elhadad, M.: Information fusion in the context of multi-document summarization. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics: College Park, MD, 20–26 June, pp. 550–557 (1999)
11. Mani, I., Barbara, G., and Eric, B. Improving summaries by revising them. In Proceedings of the 37<sup>th</sup> Annual Meeting of the Association for Computational Linguistics: College Park, MD, June, pp. 558–565 (1999)
12. Lin, C. Improving Summarization Performance by Sentence Compression- A Pilot Study.In the Proceedings of the Sixth International Workshop on Information Retrieval with Asian Language (IRAL): Sapporo, Japan, July 7 (2003)

13. Knight., Marcu, D.: Statistics-Based Summarization- Step One: Sentence Compression. In Proceedings of AAAI: Austin, TX, USA (2000)
14. Hovy, E., Lin, Z. L.: A BE-based Multi-document summarizer with sentence compression. In Proceedings of Multilingual Summariza-tion Evaluation (ACL), Ann Arbor, MI (2005)
15. Liu, H.: MontyLingua: An end-to-end natural language processor with common sense.: Available at: [web.media.mit.edu/~hugo/montylingua](http://web.media.mit.edu/~hugo/montylingua) (2004)
16. Dorr, B. Zajic, J., David, S. R.: Hedge trimmer: A parse-and-trim approach to headline generation. In Proceedings of the HLT/NAACL Text Summarization Workshop and Document Understand ing Conference (DUC): (pp. 1–8). Edmonton, Alberta (2003)
17. Hovy, E.H., Fukumoto, J., Lin, C.-Y., Zhou L.: Basic Elements.: <http://www.isi.edu/~cyl/BE> (2005)
18. Barzilay, R., Elhadad., McKeown, K.: Sentence ordering in multi-document summarization. In Proceedings of the Human Language Technology Conference. (2001)
19. Lin, C.-Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence. In Proceedings of Language Technology Conference (HLT-NAACL);, Edmonton, Canada, May 27 - June 1 (2003)
20. Lin. C.Y.: ROUGE: A package for automatic evaluation of summaries. In WAS 2004: Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain July 25–26 (2004)

# **STAIR: A System for Topical and Aggregated Information Retrieval**

C.V. Krishnakumar<sup>1</sup> and Krishnan Ramanathan<sup>2</sup>

<sup>1</sup>Stanford University, California, USA

<sup>2</sup>HP Laboratories , India

**Abstract:** Web content has exploded dramatically in the last decade and search is becoming increasingly complex. In the current search paradigm, the user has to enter the query and is immediately presented results that are typically accessed sequentially. However, there are scenarios where the above model is not appropriate, either because results being in consumable form is more important than immediacy of results, or because the it is difficult and time consuming to navigate the results in sequential fashion. In this work, we describe the architecture, implementation and utility of STAIR- The System for Topical and Aggregated Information Retrieval, that uses a variant of focused crawling and retrieves just the relevant information from the web. We present a new interface that selects search results from different search engines, ranks the results and presents the most relevant results as an aggregated PDF document. User studies indicate that the relevance of the results produced by our approach is competitive with those of current search engines

## **1 Introduction**

Search engine technology has had to scale dramatically to keep up with the growth of the web. The one-size-fits-all approach that is being used by the general purpose search engines today is increasingly becoming irrelevant today. Search interfaces today are geared to providing results immediately and getting users to click relevant ads. The results are presented as a sequence of links and snippets from the linked-to document.

The need is for a system that would provide an all round approach that could provide the most "relevant" information about the given query within acceptable time limits. Our solution is to design a information assistant that queries multiple search engines based the information need, selects and consolidates the results and presents them to the user in a compact and consumable manner. The response is provided as a PDF document containing multiple articles. Navigating the consolidated document is much simpler, the user gets more information (compared to search result snippets) that enables her to quickly decide whether to

read the content or move to the next result. We believe such an interface would be even better suited for newer kinds of internet access points such as \* In mobile devices, where there is a chance of losing internet connectivity while navigating search results (because the user is on the move and the connection drops) and where it is more cumbersome to surf through multiple results on multiple web pages. \* Touch based devices where the traditional keyboard/mouse based interaction does not provide as good an experience as our new interface can. In this work, we present the prototype of *STAIR* - System for Topical and Aggregated Information Retrieval that implements focused crawling and retrieves only the relevant information from the web. It compiles ,consolidates and processes the information to provide an aggregated PDF document.

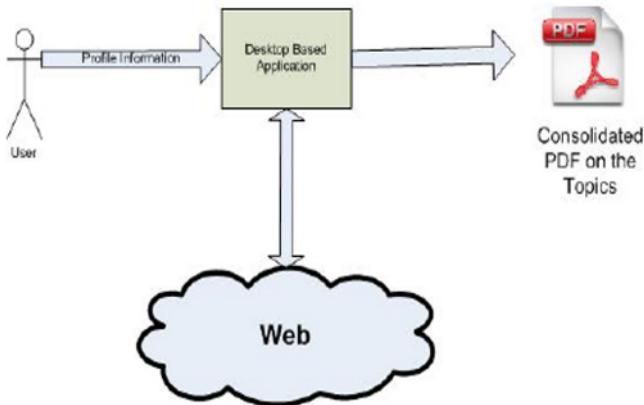
## 2 Problem Definition

To create an *aggregated* and *personalized Information Retrieval* (IR) system that **compiles** and **consolidates** the most relevant information on particular topic(s) from the web and automatically creates a *document* providing comprehensive info on topic. The information is provided in terms of facts extracted from the web. This projects thus attempts to *personalize the user experience on the web* and enable an on-the-fly retrieval and archival of the user specified content from the web.

### 2.1 Deployment

The system can be deployed in primarily two possible ways:

The system can be deployed as a desktop-based application wherein, in conjugation with the User profile generator, it would take the input as the user profile and generated the *consolidated magazine* for each of the topics the user is interested in [Fig 1].



**Fig.1.** The System as a black box

The system can also be deployed as a *Web Service* wherein, the system would reside on the Server, the user would give his topics of interest and the system would generate the *consolidated magazine* based on the current relevant information extracted from the web.

### 3 Motivation

In recent years, the World Wide Web has grown at a rapid pace. [1] In many cases, it is desirable for the user to find material on the topic of his interest. In order to achieve this goal, there are a number of search engines that facilitate this goal. However, different search engines have different coverage of the web. Although meta-search engines have been tried in the past, they have been unsuccessful largely because they are unable to scale on the server side. Moreover, there is a tradeoff between getting results immediately and getting relevant results. The user may be willing to wait for some time (e.g. 10 minutes) if the search engine could do a better job of filtering the results. There is no interface for specifying user wait time today. However, the conventional search engines attempt to serve the public as a whole and thus try to collect and index all the documents on the web. There are some drawbacks of this kind of approach:

1 . We need a mechanism whereby we can hope to retrieve and store only the most relevant pages from the web. To achieve this end, we use a focused crawling method, so as to retrieve and index only the most relevant pages from the web. In this manner, we also hope to leverage the knowledge of the user specified topics beforehand.

2. The most popular web search engines of today use a *one-size-fits-all* approach in order to serve the entire world. Since the queries are user specific, this approach also succeeds in personalizing the web for each user.

3. The current web search engines have navigational aims besides information extraction. Our project does not have the navigational aims. On the other hand, our system attempts to satisfy just the informational needs of the user. Thus, the system provides the compiled and consolidated information on the user specified topics, into a magazine, that can be archived. Drawing an analogy, whilst the web search engines act a *news ticker*, our system provides a *consolidated magazine* in the pdf format, perfectly suitable for archiving.

## 4 Related Work

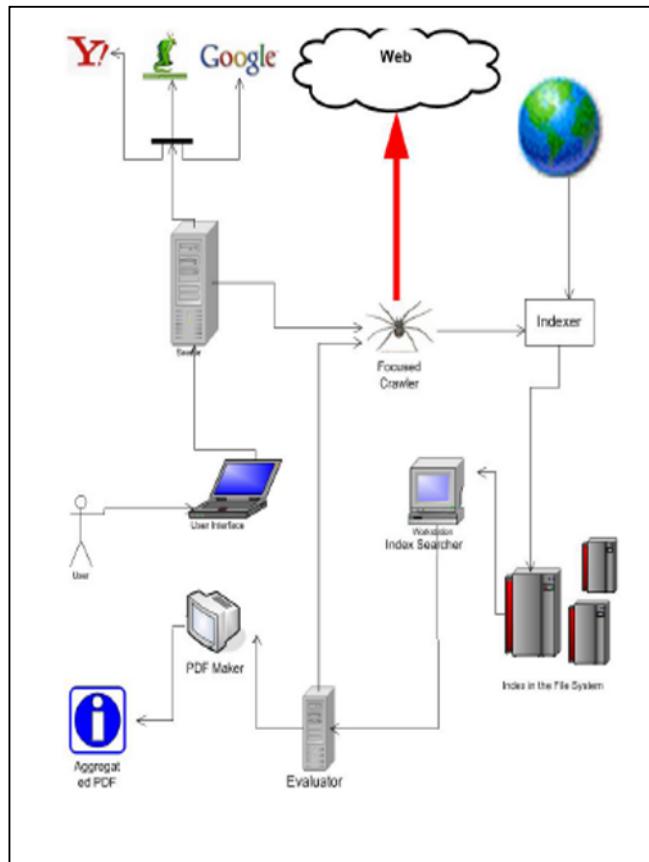
Many different approaches have been proposed by the researchers in the recent years to achieve the goal of improving the efficiency and the accuracy of the search engines, by avoiding the fetching of irrelevant pages from the web. The pioneer *focused crawler* introduced by Soumen Chakrabarti[9] used a topical taxonomy and graph distillation to track topical hubs. The Volant system provides a information retrieval paradigm taking post-query navigation into account. White et.al propose a machine learning-based approach for supporting switching search engines by estimating in real time whether more accurate results exist on alternate search engines. The Clusty search engine from Vivisimo clusters results and presents them using the desktop metaphor of folders. To overcome the low recall problems, Bergmark suggested Tunnelling[?], an attempt to go through the low relevance pages to the highly relevant pages. Chakrabarti also suggested resource discovery through examples [3]. One other important work in this field has been the combination of link and text analysis for focused crawling by Almanidis. The hyperlink features for personalization such as URLs, tokens and anchor texts were suggested by Aktas [11]. Perhaps, the work closest to our system is BINGO [2] which provides an architecture for focused crawling. However, the overall emphasis in *BINGO* is on focused crawling unlike our system which places an equal emphasis on all aspects of Information Retrieval such as Fetching, Query Expansion, Focused Crawling, Ranking and Presentation.

## 5 Overview and Architecture

We propose the following architecture for *STAIR* system. The main components of the system are depicted in the Fig. 3. The major components include:

- Seeder
- Fetcher
- Focused Crawler
- Indexer
- Index Searcher

- Evaluator and Ranker
- PDF Maker
- User Interface



**Fig. 2.** System Architecture

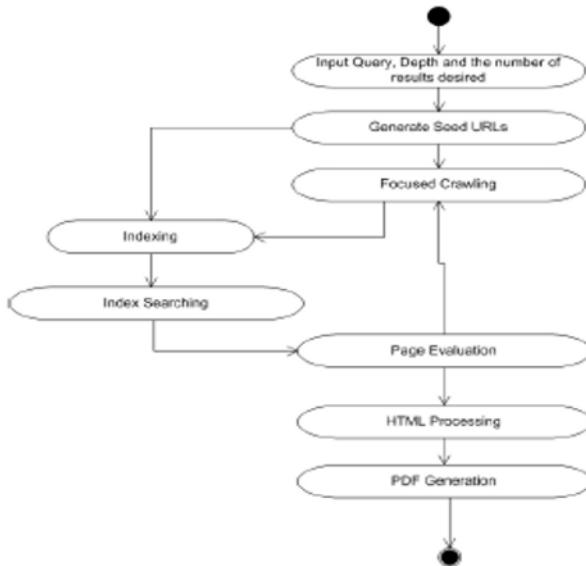


Fig.3. Overall Flowchart

## 6 Design and Implementation

### 6.1 Implementation

The entire system was implemented in Java. The system is Platform independent and highly portable. It is currently implemented as a *Desktop based live , 'on-the-fly search' search and pdf production mechanism* and can be easily extended into a web service by incorporating it as a servlet.

### 6.2 Modules

The system is highly modular with distinct modules placed into clearly demarcated packages. The modules, the packages and their primary functionalities are described in this section.

### 6.2.1 Input Processor

This module is the first module of the system *STAIR*. It is the module that acts as a plug-in to the user profile. It also has the ability to take the input in the form of a user profile, which, typically is in the following format:

tendulkar->10.0 data mining->25.0

Such an input is stored in a map-like structure and used by the system for the information extraction and retrieval.

### 6.2.2 Seeder

The Seeder module is a very important component of *STAIR* since it provides the system with the initial set of URLs for the information retrieval. The initial set of URLs are currently being obtained from the General purpose search engines, viz. Google, Yahoo! and DMOZ. The seeder also performs the function of converting the raw HTML web-pages into an organized structure. This organized structure is termed as the *SearchResultNode* and provides the metadata about the page in addition to the page itself.

### 6.2.3 Crawler

The crawler module is the most important module in our system, since it distinguishes our system from the general purpose search engines. Traditional search engines such as Yahoo! and Google use a *Breadth-First-Search Approach* for retrieving the information from the web. On the other hand we use a modified version of the the breadth-first-search known as the *best-first-search* for selective topic extraction.

This approach of selectively fetching only few of the pages on the frontier list is only known as *focused crawling*. Proposed by Soumen Chakrabarti [9][1], it has been used in some systems, the most recent version being in [5].The primary advantage of focused crawling is that the space complexity is minimum since only the most relevant links are fetched and indexed at each iteration. However, this comes at the cost of time complexity since, the system should now combine the evaluation phase with the crawling phase to predict the most rewarding set of links beforehand. This has been implemented by heuristics in the current version of *STAIR* wherein we predict the score of the child link as being times that of the parent node and then consider just 25% of the frontier List at each stage.

### 6.2.4 Indexer

Indexing is the process of converting the documents into a format that is very easily searchable. The indexer is the module that is responsible for indexing the web pages into the file system. We use a Lucene based index The synonym Analyzer we use also enables our system to extrapolate the search to *Semantic*

dimensions rather than mere text matching. The token analyzer and the engine are highly modular and can easily be extended to cover concept matching too.

### 6.2.5 Searcher

Searcher is the module that performs the search on the index and returns the results that satisfy the query. The searching takes place through the IndexSearcher() provided by Lucene. It returns a list of 'HITS' objects. The HITS objects that are returned are ranked by their score. The scoring of *STAIR* system is explained in the next subsection.

### 6.2.6 Evaluator

This subsection assigns relevancy scores to each document in the index. This score is extremely important since it plays a vital role in the *frontier list* selection and consequently in adding focus to the crawler.

There are different parameters on which the documents are evaluated. They are described as follows: **[Content]**: The content is used in the scoring by considering the product of Term frequency and the inverse document frequency. The term frequency in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents which may have a higher term frequency regardless of the actual importance of that term in the document to give a measure of the importance of the term within the particular document .

**[Collaborative Filtering]**: This feature is an unique feature of our system. We leverage the power of collaborative effort to estimate the document similarity. Instead of relying on the Vector spaces of the document content,we model the document as a vector space of its Delicious Tags.The advantages of modeling the document on its Delicious Vector is that instead of relying on the textual content to describe the purpose of the document, we rely on the tags that have been assigned to the document by the human users around the globe, depicting its primary goal.We use the concept of *Cosine Similarity* to find the similarity between the web pages. For instance, the similarity between

www.gmail.com

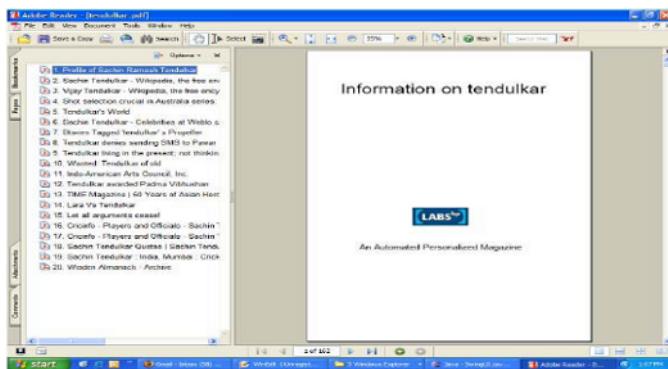
and

www.mail.yahoo.com

was found to be reasonably close to 1 even though the textual similarity would have yielded a much lesser number.

### 6.2.7 Presenter

The presenter module is responsible for the HTML processing to a human readable form and subsequent PDF generation for archival.



**Fig. 4.** First Page of the PDF Output

### 6.2.8 User Interface

The User Interface module provides a compact and a highly functional interface to the *STAIR* system as shown. The User Interface has been built using Swing, a 100% pure java interface. The User interface has the inputs:

- Search Query
- Depth of Crawl
- Number of Results

The *Search query* represents the topic of interest for the User. The *depth of crawl* depicts the number of iterations that need to be performed before the system terminates. The *Number of Results* represents the number of results desired in the PDF document. The output from the system is a PDF document that can be saved on to the system by the user.

## 7 Evaluation

### 7.1 Conventional Parameters of Evaluation

The traditional methods of evaluation of Search Algorithms revolve around the concepts of *Precision* and *Recall*. Precision measures the proportion of documents in the result set that are actually relevant. Recall measures the proportion of all relevant documents in the collection that are in the result set.

### 7.2 User Study

Relevance is hard to measure on the web. However, the level of user satisfaction with a search service can be taken as a rough measure of the relevance

of the results provided. To evaluate our system, we conducted an User study so as to draw a comparison between the kinds of results given by our system as against those given by the general purpose search engines, viz. Yahoo! [<http://search.yahoo.com>] and Google [<http://www.google.com>]. The system does not aim to replace the actual general purpose web-search engines. The experiment was carried out only to determine how relevant the users perceived the results produced by our system against those generated by the state-of-the-art search engines. The users were asked to rank the resultant URLs on a scale from 0 to 10 without revealing the source or the ranking of the URLs. Some sample results, showing the perceived relevance at 10, is summarized in the graphs that follow.

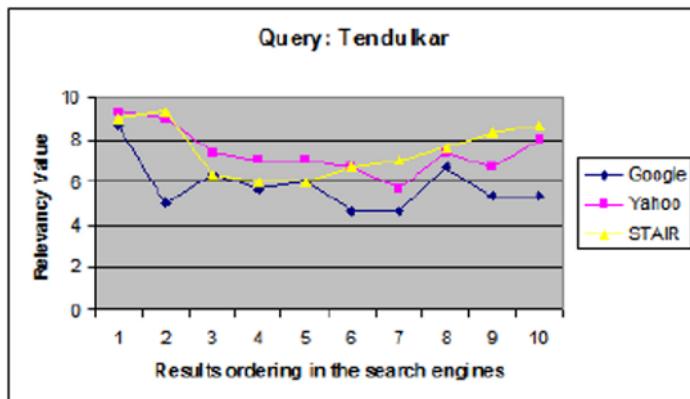


Fig. 5. Results for the evaluation for the query: Tendulkar

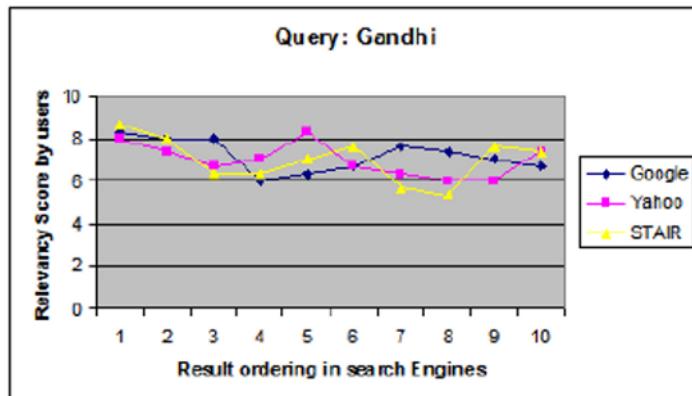


Fig. 6. Results for the evaluation for the query: Gandhi

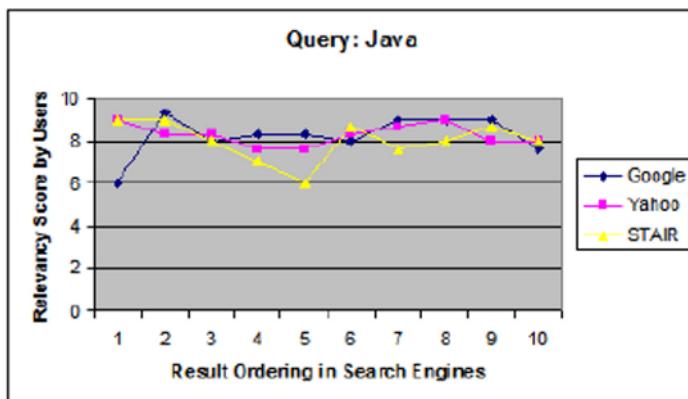


Fig. 7. Results for the evaluation for the query: Java

### 7.3 Inferences

The graphs show that the quality of the links produced by our system is comparable to those by the general purpose search engines. Since the space complexity of our system is much lesser than that of the General purpose search engines, due to the *focused crawling approach* used. In addition, our system consolidates and provides the links as well as the content as a document that can be stored, archived and referred to at a later date.

## 8 Future Scope

The current system is a desktop based system that takes as its input an user query and gives as its output the consolidated pdf on a particular topic.

: The quality and the readability of the output documents can be enhanced by using more powerful text extraction and noise removal techniques through machine learning.

: The browser history and bookmarks provide a valuable mine of information regarding the user behavior. This information will be mined effectively to build a user profile that could be given as an input to this system.

: Currently, the *Wordnet* is used as the only ontology for extracting and placing synonyms of a word in the index. We plan to move one step ahead towards Semantics based search by using the power of the *Wikiconcepts* in the Analysis stage. The current system is easily extendable to any such plug-in in the future.

### 8.1 Limitations of the System

: Since the entire system has been implemented on a desktop based system and since currently the queries are processed on the fly, the computational time complexity of the system is much higher than that of the web search engines.

: The current system also produces some noisy data in the resultant pdf in the form of the text of the sidebars. This noise has been localized currently by the use of sentence extraction techniques. However, this problem can be resolved by the use of more powerful text processing mechanisms, so as to provide a perfect output.

## 9 Conclusion

In this work, we have present a prototype of a robust and a platform independent system *STAIR* for focused Crawling, retrieval and presentation of consolidated user-specific information from the unstructured web. We also demonstrate that the focused crawling is a powerful means of reducing the overheads associated with the *one-size-fits-all approach* that the traditional search engines follow. Our system,built over Lucene , explores the web in a focused manner, guided by the relevance of the documents it finds. It filters the data at the data-acquisition level. Moreover, unlike the traditional search engines that provide just links, our system extracts, cleans and consolidates the content from the web into a PDF document, providing a novel user experience.The user study gives an indication the quality of the links provided by our system is comparable to those of the general purpose search engines.

## References

1. Aggarwal, C. C.: Learning Strategies for Topic Specific Web Crawling. IBM T.J. Watson Research Center.
2. Sizov, S., Biwer M., et al.: The BINGO! System for Information portal Generation and Expert Web Search. CIDR Conference. (2003)
3. Chakrabarti, S., Martin, H., Berg, V. D., Dom B.: Distributed Hypertext Resource Discovery through Examples. VLDB Conference. (1999)
4. Hersovici M., Jacovi M., et al.: Shark Search Algorithm. An application : Tailored Web Site Mapping, Elsevier. (1998)
5. Pandey S., Olston C.: Crawl Ordering by Search Impact. WSDM '08. (2008)
6. Raghavan P. et al.: Introduction to Information Retrieval. Cambridge University Press. (2008)
7. Pandit S., Olston C.: Navigation aided retrieval, WWW conference, 2007. pp. 391-400. (2007)
8. Ryen White et.al.: Enhancing web search by promoting multiple search engine use, SIGIR 2008, pp.43-50.(2008)
9. McBryan O. A.: Tools for Taming the Web, First International Conference on the World Wide Web. GENVL and WWW CERN. Geneva (Switzerland). May 25-27 (1994)

10. Brin S., Page L.: The anatomy of a large-scale hyper textual web search engine. WWW7. pp 107 – 117. (1998)
11. Aktas M. S., Nacar M. A., Menczer F.: Using Hyperlink Features to Personalize Web Search. Indiana University.

# **Exploring Multiple Ontologies and WordNet Framework to Expand Query for Question Answering System**

Santosh Kumar Ray<sup>1</sup>, Shailendra Singh<sup>2</sup> and B. P. Joshi<sup>3</sup>

<sup>1</sup>Birla Institute of Technology, Muscat, Oman, santosh@waljatcolleges.edu.om,

<sup>2</sup>Samsung India Software Center, Noida, India, shailendra.s@samsung.com,

<sup>3</sup>Birla Institute of Technology, Noida India, bp\_joshi@yahoo.com

**Abstract.** Query expansion plays a very important role in enhancing the performance of the Question Answering Systems. There are several methods proposed for query expansion and the use of ontologies has been the latest and popular choice of the researchers because of its effectiveness in building conceptual query. However most of these query expansion methods have used either a single domain specific ontology or WordNet and none of research work has been reported the use of ontologies and WordNet together. In the context of Worldwide Web based Question Answering Systems, the use of single ontology or WordNet has not proved to be sufficient to retrieve wide variety of heterogeneous information. In this paper, we have proposed an efficient query expansion method that uses multiple ontologies retrieved from semantic web search engine such as Swoogle and combines them with WordNet to disambiguate the context. We have experimented on a set of 300 questions collected from TREC and other resources to judge the accuracy of the proposed method. We have shown results using Google as well as with respect to few existing popular web-based Question Answering Systems like START, AnswerBus, BrainBoost, and Inferret.

## **1 Introduction**

Today, World Wide Web has become the chief source of information for everyone from general user to experts, students to researchers, to fulfill their domain specific needs. Search engines like Google help users to find the relevant information based on the keyword searching and retrieve a large number of links. However, in many cases none of retrieved web pages contain the relevant answer. In last few years, Question Answering Systems have emerged as a good alternative to provide relevant answers of user queries in succinct form.

A typical Question Answering System takes user's question in some natural language as an input. This question is then optionally modified using some query

modification technique (also called query expansion) and output of this modification process is a set of queries similar in meaning to original question. These questions are fed into knowledge repository which may be either predefined corpus (as in case of START [13]) or entire World Wide Web (as in case of AnswerBus [3] or Inferret [9]). Documents containing answers of the user query or modified queries are retrieved from the knowledge

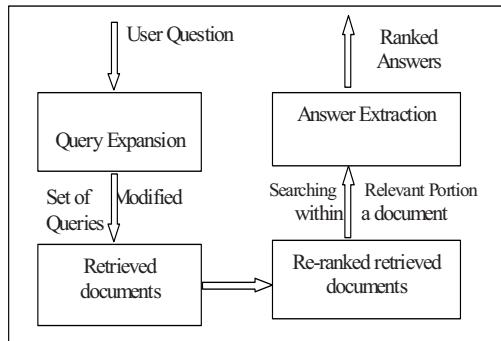


Fig. 1. A typical Question Answering System Architecture

repository and re-ranked based on their relevance to the user query. Finally the most relevant portions of the predefined number of documents along with links are presented as answers to the user's question. As shown in figure 1, a typical Question Answering System consists of four major phases. Though all of these phases are equally important but query expansion is having major role to boost up the recall of the Question Answering Systems. In this paper we are focusing on query expansion to add related concepts in query processing.

For web-based open domain Question Answering Systems, one or two ontologies are not sufficient to identify the correct sense of words. Hence we are using multiple ontologies and combining them with WordNet [15] to disambiguate and identify the correct sense of the concepts in user query. At present, it is very difficult to find the suitable set of ontologies. There are few semantic web search engines available on WorldWideWeb such as Swoogle [6], OntoSearch [16], AKTiveRank [2], OntoClean [8], and OntoKhoj [11] which are maintaining repositories for a large number of domain ontologies. We are using Swoogle as Ontology database for our query expansion method because Swoogle is having the largest number of ontologies and updates its ontology base periodically.

In this paper, section 2 provides details of related research work in the field of query expansion using ontologies and WordNet. Section 3 explains the proposed method for query expansion using multiple ontologies and WordNet. We have shown results of our experiments in section 4. In the last section, we have stated conclusion and future direction to build an intelligent Question Answering System.

## 2 Related Work

Use of ontologies for query expansion has become popular in recent years. However more focus has been given on the use of single domain ontology while use of multiple ontologies is quite rare. [12] has combined Web ontology and WordNet together. However focus of their work is to create web document representation rather than query expansion. Further they have used a single ontology and restricted the number of semantic relations to perform their experiment. In [10] Google has been combined with WordNet for word sense disambiguation. However, they have restricted their experiment to noun terms only. Also aim of word sense disambiguation in their project is ontology learning and not for query expansion. They have shown a small improvement in word sense disambiguation by combining WordNet with Google. [5] has used multiple ontologies for query expansion. However, their experiments are using limited number of ontologies. In the contrast, our proposed approach is using multiple ontologies accessed from Swoogle which dynamically includes the ever increasing ontologies on the semantic web. Further, they have not included WordNet for their query expansion method. CIRI system [1] visualizes the ontologies as subsumption trees, from which concepts can be selected to constrain the search. The actual search is done through keywords annotated to these concepts and sub-concepts, using a traditional search engine. However in this system user selects the relevant concept in ontology. Our proposed system is similar to system discussed in [7] that uses ontologies pool obtained from Swoogle and other lexical resources such as WordNet. The aim of their system is to find all possible keyword sense using ontology pool. On the other hand, our system aims to find only those senses of keywords that are closer to the domain of the other keywords existing in the user question. Keeping in view the time complexity and large resources provided by Swoogle, we have limited ourselves to Swoogle for searching of ontologies.

## 3 Multiple Ontologies and WordNet-based Query Expansion Methodology

Ontologies and WordNet are having rich information about domains and semantic relations between concepts. Query expansion methods based on only selected relations in WordNet has resulted into degradation of Question Answering Systems performance while on the other hand, if we use all the relations in WordNet in an uncontrolled manner then we get more number of semantically related words which forms large number of modified queries against user's single question. So, this method is not viable because of more computational resources.

We are proposing multiple ontologies and WordNet based query expansion method. The proposed method takes user question as an input and extracts key concepts from the question then automatically finds the most relevant senses for key concepts of the question using WordNet. To disambiguate the correct sense,

the method uses multiple domain ontologies retrieved from Swoogle search engine. When key concept(s) from the user's question is fed into Swoogle then it returns ontology classes describing the concept(s). In the proposed method, we compute semantic distance of the key concepts from retrieved ontology class, super class, and its subclasses and consider the class with the lowest semantic distance for the further processing. The complete algorithm is given as follows:

**Algorithm:** Query\_Expansion\_MultipleOntologies

**Input:** User's Question considered as Query (Q)

**Output:** Expanded query ( $Q_E$ )

**Step 1:** Let  $T$  be a set of quadruples and defined as  $T = \langle C, O, W, R \rangle$ , where  $C$  denotes concept in user's question,  $O$  represents ontology for the concept  $C$ ,  $W$  represents weight of an ontology  $O$ , and  $R$  represents one of the semantic relations retrieved from WordNet. Initially,  $T$  is empty.

**Step 2:** User enters a query  $Q$ .

**Step 3:** Extract key concepts  $C_1, C_2 \dots C_k$  from  $Q$ .

**Step 4:** User assigns  $W_1, W_2 \dots W_k$  weights to the concepts  $C_1, C_2 \dots C_k$  on the scale of 1-10. The concepts with higher weights are considered as important concepts.

**Step 5:** Search Swoogle for the combination of concepts using term dropping strategy. Query for the Swoogle is fed into Conjunctive Normal Form. All ontologies describing a concept combination are put into one group. Let us assume ' $n$ ' ontology groups defined as  $OG_1, OG_2, \dots, OG_n$ .

**Step 6:** Let  $WN_{c1}, WN_{c2}, \dots, WN_{ck}$  be the domain set in WordNet for concepts  $C_1, C_2 \dots C_k$ . Elements in  $WN_{ci}$  are denoted by couple  $(S, R)$  where  $S$  is synonymous set for concept  $C_i$  and  $R$  is relation in WordNet that connects  $C_i$  to  $S$ .

**Step 7:** for  $(i=n; i>0; i--)$  do following for each ontology of  $OG_i$  group.

$T = T \cup (\langle C_1, x, W_{OGi}, R_1 \rangle, \forall x \in (O_{ij} \cap WN_{c1}) ) \cup (\langle C_2, x, W_{OGi}, R_2 \rangle, \forall x \in (O_{ij} \cap WN_{c2}) ) \cup \dots \cup (\langle C_k, x, W_{OGi}, R_k \rangle, \forall x \in (O_{ij} \cap WN_{ck}) )$  (where  $O_{ij}$  is  $j$ th ontologies of ontology group  $OG_i$ )

**Step 8:** If  $T$  is empty

for each  $C_i$

add one sense from all relations available in WordNet to  $T$ . We select most frequently occurring sense of the word and assign zero weight to the ontologies.

**Step 9:**  $Q_E = (C_1 \text{ OR } O_{11} \text{ OR } O_{12} \dots \text{ OR } O_{1m}) \text{ AND } (C_2 \text{ OR } O_{21} \text{ OR } O_{22} \dots \text{ OR } O_{2n}) \text{ AND } \dots \dots (C_k \text{ OR } O_{k1} \text{ OR } O_{k2} \dots \text{ OR } O_{kr})$  where  $O_{ij}$  is the common ontology for concept  $C_i$  found in previous steps.

We explain discussed algorithm with the help of an example. Let us consider a query in the form of question i.e. "What is Jupiter's atmosphere made of?". The key concepts found in this question are written as  $C_1 = \text{"Jupiter"}$ ,  $C_2 = \text{"atmosphere"}$ , and  $C_3 = \text{"made"}$ . User assigns weights to each concept as  $W_1 = 9$ ,

$W_2=9$ ,  $W_3=3$  respectively. We do start searching of relevant ontologies from Swoogle using term dropping strategy. Swoogle provides many advanced meta-tags for specific search. We are using meta-tag called “*desc: term1*” to retrieve ontologies which are having “*term1*” in the description of the document, generally in the annotations. We are passing queries to Swoogle in Conjunctive Normal Form to retrieve ontologies which are relating given concepts in some meaningful way. Swoogle uses “AND” as a default logical operator. We present retrieved ontologies in table 1.

In step 7, we find common concepts between ontology group and WordNet group like in 5<sup>th</sup> query, concept *planet* is common in both OG<sub>5</sub> and WN<sub>Jupiter</sub>. Hence quadruple for this query will be <Jupiter, Planet, 9, hypernym> and will be added in T. We do step 7 and step 8 for all queries and get final set T = {<atmosphere, air, 18, synonym> <Jupiter, Planetary Object, 18, hypernym>, <Jupiter, Planet, 9, hypernym> <atmosphere, weather, 9, hypernym>, <make, constitute, 9, synonym>}. Therefore, final expanded query is defined as Q<sub>E</sub> = [(Jupiter OR Planet OR Planetary Object) AND (atmosphere OR air OR weather) AND (make OR constitute)].

As described in step 6, we present semantically related concepts for C<sub>1</sub>, C<sub>2</sub>, and C<sub>3</sub> retrieved from WordNet as in table 2. In the next section, we will run our proposed algorithm for large number of questions and will also compare results with existing Question Answering System.

**Table 1.** Ontologies groups retrieved from Swoogle for the query “*What is Jupiter's atmosphere made of?*”

N	Queries fed into Swoogle	No. of retrieved Ontologies	Listing of Top 10 ontologies (OG)
	<i>desc:Jupiter</i> <i>desc:atmosphere</i> <i>desc:make</i>	0	OG <sub>1</sub> : {}
	<i>desc:Jupiter</i> AND ( <i>desc:atmosphere</i> OR <i>desc:make</i> )	4	OG <sub>2</sub> : {Day, Day_5, Day_5, Day7}
	( <i>desc:Jupiter</i> OR <i>desc:make</i> ) AND <i>desc:atmosphere</i>	13	OG <sub>3</sub> : {Air, Air, BallisticMissile, PlanetarySurfaceObject, GravitationallyBoundObject, PlanetarySurfaceRegion, BallisticMissile, CloudFormation, SurfaceToSurfaceMissile-Ballistic, BallisticMissile}

	$(desc:Jupiter \text{ OR } desc:atmosphere) \text{ AND } desc:made$	8	$OG_4: \{PhysicalSubstance, FootballBall, olfaction, Mesh\_D\_01.268.150.250, Mesh\_D\_01.268.150.250, SPACE\_PROBE, SPACE\_PROBE, space\_Probe\}$
	$desc:Jupiter$	67	$OG_5: \{Origin, Planet, Asteroid, PerceiveThings, Magnetosphere, Day, Region, Jupiter, Observed region\}$
	$desc: atmosphere$	855	$OG_6: \{Continuant, IndependentContinuant, Site, IndependentContinuant, Atmospheric\_Phenomenon, Weather, Continuant, Shooting\_Star, Oxygen, Outer\_space\}$
	$desc: made$	16184	$OG_7: \{maker, Rights,Made, Has\_Creator, Forum,Provenance, focalLength, Concept\_Scheme, Artifact, Artefact\}$

**Table 2.** Semantically related concepts in WordNet for the query “What is Jupiter’s atmosphere made of?”

$WN_{Jupiter} = \{(“Solar System”, holonym), (“planet”, hypernym), (“Roman God”, domain Category)\}$
$WN_{atmosphere} = \{(“ambiance”, synonym), (“air”, synonym), (“Aura”, synonym), (“genius loci”, hyponym), (“condition”, hypernym), (“Standard Temperature and Pressure”, Hyponym), (“pressure unit”, Hepernym), (“air space”, hyponym), (“ionosphere, meronym), (“region”, hypernym), (“Earth world globe”, holonym), (“sky”, hyponym), (“exosphere, mesosphere, stratosphere, thermosphere, troposphere”, meronym), (“gas”, hypernym), (“mystique”, hyponym) (“weather”, hypernym)\}$
$WN_{make} = \{(make, produce , create etc...not full list)\}$

## 4 Results

To judge the accuracy of proposed query expansion method, we have considered a set of 300 questions collected from TREC [14] which are covering almost 30 different domains. In table 3, we are representing top 20 questions and their corresponding expanded queries. We fed 300 original questions and their expanded queries in Google separately and evaluated top 10 retrieved results for each original question as well as for each expanded query. With the set of original

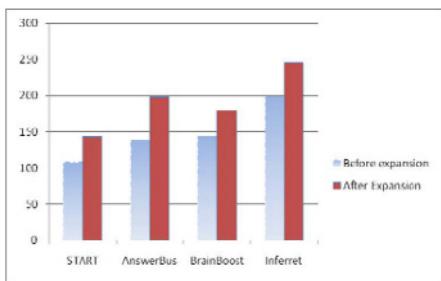
questions, we find satisfactory answers for 258 out of 300 questions while with the set of expanded queries satisfactory answers retrieved for 276 out of 300 questions. The proposed approach has retrieved 92% answers correctly and an overall improvement of about 8% in comparison with answers retrieved for existing set of original questions.

**Table 3.** Listing top 20 original questions and their corresponding expanded queries based on multiple ontologies and WordNet

1	What is Jupiter's atmosphere made of?	(Jupiter OR Planet OR Planetary Object) AND (atmosphere OR air OR weather) AND ( make OR constitute)
2	Explain the reason for sky's blue color.	( reason OR cause ) ( sky OR rainbow OR cloud OR lightning) AND ( blue OR sky-blue) AND color
3	Which planet has the least surface area?	planet AND ( least OR smallest OR minimum) AND surface AND area
4	What capital is on the Susquehanna River?	(capital OR “state capital” OR means OR Centre OR “Graphic symbol”) AND Susquehanna AND river
5	How far is Mars from our planet?	(far OR distant) AND (Mars OR “Red Planet”) AND our AND (Planet OR “terrestrial planet”)
6	What is famous invention by Marconi?	(famous OR celebrated OR known OR notable) AND ( invention OR creativity OR creativeness OR “creative thinking” OR “ creating by mental act”) AND ( Marconi OR Guglielmo Marconi)
7	What is the life expectancy of the average woman in Nigeria?	(“Life Expectancy” OR “Life Expectancy at Birth” ) AND average AND woman AND ( Nigeria OR Lagos OR Zaria OR “Yerwa Maiduguri” OR Niger OR Africa )
8	Give me the countries that border India.	( Countries OR country OR land) AND ( border OR “has border” OR “borders on”) AND( India OR Indian)
9	Which is the deepest sea?	( deepest OR deep) AND ( sea OR Ocean)
10	What continent is India in?	(continent OR subcontinent OR landmass OR Asian OR African) AND ( India OR Indian)
11	Which state has the longest coastline on the Atlantic Ocean?	(state OR province) AND ( longest OR long OR length) AND coastline AND Atlantic AND ( Ocean OR “Atlantic Ocean” OR “ Pacific Ocean” )
12	What fraction of the ozone layer is destroyed?	Fraction AND (Ozone OR “Ozone layer” OR stratosphere Or oxygen) AND layer AND depleted
13	What year did Beethoven born?	Year AND ( Beethoven OR Ludwig van Beethoven OR music OR composer) AND( born OR “born in”)

14	Who is the composer of opera semiramide?	Composer And ( opera OR “comic opera” OR “opera bouffe” OR bouffe OR “opera comique” ) AND semiramide
15	What music did Debussy compose?	(music OR Bach) AND (Debussy OR Claude Debussy, Claude Achille Debussy) AND (composed OR compose OR composer OR “composed for ” OR “ is composed of” OR “ composed for” OR “ music composed by”)
16	In what year did Arundhati Roy receive a Booker Prize?	Year AND (“Arundhati Roy” OR Arundhati) AND ( get OR receive) AND Booker AND( prize OR award)
17	Who was the seventh president of India?	(Seventh OR 7 <sup>th</sup> ) AND president AND (India OR Indian)
18	What Indian state has the highest life expectancy?	Indian AND (state OR province) AND highest AND (“life expectancy” OR “ life expectancy at birth”)
19	What does Taiwan flag look like?	(Taiwan OR Taiwanese OR Taipei OR “South China sea”) AND ( flag OR “national flag”) AND ( look OR appear) AND (like OR “likes of”)
20	What famous communist leader died in Mexico City?	Famous AND communist AND leader AND (died OR “died in year” OR death) AND Mexico AND ( city OR Leon OR “Acapulco de Juarez” OR Tepic OR Culiacan OR Matamoros OR “Tuxtla Gutierrez” )

To measure the performance of proposed approach, we have experimented with some existing popular web-based automatic Question Answering Systems like START, AnswerBus, BrainBoost, and Inferret. We reformulated expanded queries as per the Question Answering System specific format and feed them in all Question Answering Systems. The overall performance of each Question Answering System shows a significant increase in retrieving correct answers. The performance bar chart has been shown in figure 2 where BrainBoost and Inferret indicate an improvement of 25%, START exhibits an improvement of 33% while AnswerBus records maximum improvement of 44%. The overall average improvement is 31.75% on Question Answering Systems which are already using very sophisticated information retrieval techniques to retrieve correct answers. On the basis of experimented results, we can say that proposed approach is working reasonably well.



**Fig. 2.** Performance of Questions Answering Systems w.r.t. original questions and their expanded questions

## 5 Conclusion and Future Directions

In this paper, we have presented that how semantic web and WordNet can be effectively utilized for web based Question Answering Systems. Semantic Web search engines like Swoogle are helping researchers to improve the efficiency of information retrieval systems. The proposed query expansion approach takes multiple ontologies and WordNet into consideration to include probable related domain concepts along with their relations. We can conclude on the basis of experimented results that combination of semantic web with vast and exhaustive lexical resources like WordNet can greatly improve performance of the Question Answering Systems. We have proposed it for query expansion phase. Similarly, this can be extended for other phases of Question Answering System. In future, we are intending to develop an efficient content based document ranking tool based on ontologies.

## References

1. Airio, E., Järvelin, K., Saatsi, P., Kekäläinen, J., Suomela, S.: CIRI - an ontology-based query interface for text retrieval. In Proc 11th Finnish Artificial Intelligence Conf. (2004)
2. Alani, H., Brewster, C., Shadbolt, N.: Ranking ontologies with AKTiveRank. In: 5th International Semantic Web Conference (ISWC 2006). LNCS, vol. 4273, pp. 1–15 Springer-Verlag (2006)
3. AnswerBus, Question Answering System, website :<http://answerbus.com>
4. BrainBoost, Question Answering System, website: <http://www.answers.com/bb/>
5. Dey L., Singh S., Rai R., Gupta S.: Ontology Aided Query Expansion for Retrieving Relevant Texts. In Advances in Web Intelligence, LNCS, vol. 3528, pp. 126–132. Springer, Heidelberg (2005)
6. Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V. C., Sachs, J.: Swoogle: A semantic web search and metadata engine. In Proc. 13th ACM Conf. on Information and Knowledge Management (2004)

7. Espinoza, M., Trillo, R., Gracia, J., Mena, E.: Discovering and Merging Keyword Senses using Ontology Matching. In 1st International Workshop on Ontology Matching at ISWC-2006 (2006)
8. Guarino, N., Welty, C.: An Overview of OntoClean. In Handbook on Ontologies. Springer pp. 151–172 (2004)
9. Inferret, Question Answering System, website:<http://asked.jp>
10. Ioannis P., Klapaftis, G., Manandhar S.: Google & WordNet based Word Sense Disambiguation. In Proceedings of the 22nd ICML Workshop on Learning & Extending Ontologies. Bonn, Germany (2005)
11. Patel, C., Supekar, K., Lee, Y., Park, E.: OntoKhoj: A semantic web portal for ontology searching, ranking and classification. In: Proceedings of the Workshop on Web Information and Data Management, pp 58–61 , ACM Press (2003)
12. Sabrina T., Rosni A., Enyakong, T.: Extending ontology tree using NLP techniques. Proceedings of National Conference on Research & Development in Computer Science REDECs 2001, Selangor, Malaysia, (2001)
13. START Question Answering System, website:<http://start.csail.mit.edu>.
14. Text Retrieval Conference, <http://trec.nist.gov>
15. WordNet, website:<http://wordnet.princeton.edu>
16. Zhang, Y., Vasconcelos, W., Sleeman. D.: Ontosearch: An ontology search engine. In Proc. 24<sup>th</sup> SGAI Int. Conf. on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK. (2004)

# **Disambiguation Strategies for English-Hindi Cross Language Information Retrieval System**

Sujoy Das<sup>1</sup>, Anurag Seetha<sup>2</sup>, M. Kumar<sup>3</sup> and J. L. Rana<sup>4</sup>

<sup>1</sup> Deptt. of MCA, MANIT, Bhopal, India, sujdas@gmail.com

<sup>2</sup> Computer Sc. & Applications, MCRPSV, Bhopal, India, anuragseetha@gmail.com

<sup>3</sup> Deptt. of Computer Sc. & IT, SIRT, Bhopal, India, prof.mkumar@gmail.com

<sup>4</sup> Director, RITS, Bhopal, India, jl\_rana@yahoo.com

**Abstract:** The information content of languages other than English are increasing rapidly on WWW. To access information of a language other than the native language we need Cross-Language Information Retrieval (CLIR). The approaches to CLIR can be classified into three different categories • document translation, query translation and interlingua matching. The dictionary based query translation approach has been widely used by researchers of CLIR. The translation ambiguity and target polysemy are the two major problems of dictionary based CLIR. In this paper, we have investigated part of speech and co-occurrence based disambiguation techniques for English-Hindi CLIR system.

## **1 Introduction**

The information content of languages other than English language is increasing rapidly on World Wide Web. Due to globalization the organizations may require contents that may be available on WWW in a language other than their native language. The solution to this is Cross-Language Information Retrieval (CLIR). It is defined as the retrieval of documents in a language other than the language of the request or query. The approaches to CLIR can be classified into three viz. • document translation, query translation and interlingua matching. The document translation approach requires that the entire documents in the collection are translated into the language of the user request. The approach may require enormous translation effort and will be expensive. In query translation approach the query is translated into the documents language and then monolingual retrieval is performed. The query can be translated using machine translation system, parallel texts and/or domain specific corpora, or Machine Readable Dictionary MRD. Query translation approach is popular among CLIR community because it is efficient and easily implemented for relatively short queries. Further the dictionary based query translation has been widely used in CLIR because of availability of MRD's. The performance of dictionary based approach may be

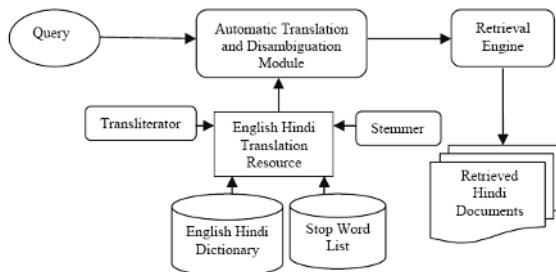
below monolingual retrieval as there is no one to one mapping from source language to target language. The failure to translate query terms and translation ambiguity are the major causes of drop in effectiveness of a dictionary based CLIR system. In this paper, we have studied the effect of target polysemy and translation ambiguity in dictionary based query translation approach for English-Hindi CLIR system. The part of speech and co-occurrence based disambiguation techniques for English-Hindi CLIR system are also studied. The Section 2 discusses related work. The Section 3 gives brief account of design of English-Hindi CLIR system and experimental set up. The Section 4 discusses various strategies for disambiguation and corresponding results are given in remaining Section. It is worth mentioning that co-occurrence based technique performed better than the part of speech based query disambiguation.

## 2 Related Work

An overview of cross language information retrieval is given in [1, 2, 3, 4]. The translation ambiguity and target polysemy are the two major problems in dictionary based CLIR. The translation ambiguity is due to the source language whereas polysemy problem occurs in target language [2]. The effectiveness of CLIR is lower than that of monolingual retrieval in simple dictionary translations [3,4]. Lisa Ballestroes and Bruce croft have developed several methods using Machine Readable Dictionary for cross language information retrieval. They have also proposed methods to disambiguate the term translation via MRD. Pirkola et. al. [5,12] have identified four major problems associated with dictionary-based CLIR, viz. phrase identification, translation ambiguity, the coverage of dictionary, and the processing of inflected word and untranslatable words. Dunning and Davis [6] have suggested parallel and aligned corpus techniques for disambiguation. David A. Hull [7] used Boolean conjunction for automatic disambiguation in the target language. Jianfeng Gao et. al. [8] used statistical models for improving query translation for English-Chinese CLIR. Fatiha Sadat et. al.[9] studied the effect of statistical query term disambiguation in Cross-Language Information Retrieval. Paul Clough and Mark Stevenson [10] have used Euro WordNet for disambiguating the Spanish queries using TREC6 CLIR test set. Mirna Adriani [11] studied the term disambiguation to select the best translation terms from the dictionary for German-French and Italian queries. Mark W. Davis and William C. Odgen [13] have used part of speech and corpus based disambiguation approach for the English-Spanish language pair. Monz and Dorr [14] have utilized bilingual dictionary and monolingual corpus for English-German language pair. Seetha, Das and Kumar have performed some experiments for the English-Hindi CLIR and the results are reported in [15].

### 3 Design of Cross Language System

The task of Cross-Language Text Retrieval remains dauntingly difficult, because of translation difficulties and moving to a new language pair may require completely new resources, and techniques[19]. The Indian language information retrieval is in nascent stage [16,17,18].The linguistic resources i.e. translation lexicons, taggers and morphological analyzers are not easily available and have posed a big challenge in the present work. The English-Hindi Cross Language Information Retrieval (CLIR) system is designed using Managing Gigabytes (MG) retrieval system. The system is divided into three modules query formulation module, query translation and disambiguation module and retrieval module (see Fig 1). The automatic query formulation module takes the input from the topics of the test collection and forms a bag of word query. The English query is then translated and disambiguated through query translation and disambiguation module. Finally Hindi documents are retrieved through MG retrieval system from the document collection which has been created by Seetha, Das and Kumar [20] for conducting English-Hindi CLIR research.



**Fig. 1.** The modules of English-Hindi CLIR System

### 4 Experimental Setup

The methodology adopted for evaluating English-Hindi CLIR system is exactly similar to the evaluation methodology of CLEF, TREC and NTCIR. Test collections are vital for empirical evaluation and comparison of the effectiveness of various approaches in the CLIR system. No standard Hindi test collection was available for CLIR research at time of inception of this research, so it was necessary to construct a Hindi Test collection as per the guidelines of TREC, CLEF. The details of Hindi test collection are given below:

**Documents:** The Hindi test collection contains 6294 Hindi documents.

**Topics:** Total 19 topics are created as per the TREC & CLEF guidelines and are used to evaluate the performance of the system. It is a known fact that average query length on web is 3 or fewer words [21]. So short queries were constructed using the title field of topic for query construction. A sample topic from Hindi test collection is shown in Table 2.0.

**Table. 1.** Sample Topic from Hindi Test Collection

```
<top>
<num> NUM19
<title> Right to Information Act
<desc> Description: Documents should contain information about Right to
Information Act
<narr> Narrative:
Relevant document should contain information about Right to Information Act. The
document may contain information about rules, provisions made under right to
information act and description about the fee.
</top>
```

**Relevance Judgment:** We have used services of language experts to provide relevance judgment for documents with respect to the topics created.

**Stop Word:** Stop words are nothing but frequently occurring non-significant words. The English stop word list was used for removing stop words from the query at the time of query formulation. The stop word list contains 507 English words.

**Stemmer:** The morphological variants are normally not present in machine readable lexicon. So a stemmer is employed to conflate the morphological variants of a word into a common root form. Stemming may help improve retrieval performance in some languages but may hurt performance in some languages [22]. The porter stemmer [23] is used for conflating the morphological variants to a stem word.

**Machine Readable Dictionary :** We have used publicly available English-Hindi machine readable dictionary (MRD) “Shabdanjali” developed by IIIT, Hyderabad [24] to translate the source English language query into the target Hindi language query. The dictionary contains nearly 25000 head words. “Shabdanjali” MRD is in ISCII format and we converted it in UTF-8 format. It further required some basic normalization.

**Transliterator:** The CLIR performance is also affected by the presence of out-of-vocabulary (OOV) words, especially named entities, newly formed words, alternate spellings, domain specific terminology, abbreviations and loan words. These words might not find entries in the lexicon due to limited coverage [25]. A significant proportion of OOV (50%) words are named entities and technical terms. Larkey et. al. [26] showed that cross language information retrieval

performance reduced more than 50% when named entities in the queries were not translated. So we used transliterator for transliterating out-of-vocabulary (OOV) words. We have developed a transliteration scheme similar to ITRANS [27].

**Part of Speech Tagger:** The Stanford part of speech tagger [28] is used for obtaining the part of speech of query term in context of the sentence.

## 5 Part of Speech Based Query Disambiguation

As explained in the Section 1, the dictionary based query translation suffers from the problem of selecting appropriate meaning from the bilingual dictionary for a query term. The main reason is that the term translation process is not isomorphic and a query term can have number of meaning in the MRD. We have tried to overcome the problem of target polysemy in source language. The aim of this strategy is to find the translation based on part of speech of a word in the query. In this strategy, we have used Stanford part of speech tagger for finding the part of speech of a word in context to given query before translating the query through machine readable dictionary. The meaning of the word has been retrieved from the MRD based on the part of speech. and the runs are referred as POS runs.

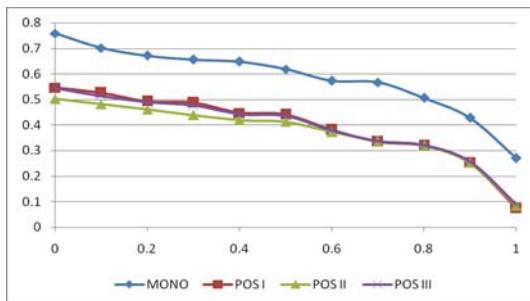
After disambiguating the query through part of speech tagger, we have used following strategies to find the impact of stop word, stemmer and transliterator.

**Strategy I (POS-I)** : Stop words are removed from the English query.

**Strategy II (POS-II)** : Stemmer and Transliterator are used on English query.

**Strategy III (POS-III)** : Stop words are removed, Stemmer and Transliterator are used on English query.

The mean average precision for the three strategies are found to be 0.3962, 0.3671 and 0.3890 respectively. A comparative recall precision graph is shown in Figure 2. We noticed that it is advantageous to remove the stop words in English-Hindi CLIR system because the performance got dropped when stop words were not removed (Strategy-II)



**Fig. 2.** Comparative recall and precision graph

## 6 Co-occurrence Based Query Disambiguation

Jianfeng [8] and Fatiha [9] have successfully used statistical techniques to overcome the problem of translation disambiguity. We have also used statistical techniques for overcoming the problem of translation ambiguity in English-Hindi CLIR system. The commonly used measures for finding co-occurrences of query terms are Mutual Information, modified dice coefficient, Log likelihood. We have used the mutual information between word pairs in the Hindi language test collection to discriminate word senses. It compares the probability of observing two events  $x$  and  $y$  together with the probabilities of observing  $x$  and  $y$  independently. If two words  $x$  and  $y$ , have probabilities  $P(x)$  and  $P(y)$ , then their mutual information,  $I(x,y)$ , is defined as:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x) \times P(y)} = \log_2 \frac{P(x/y)}{P(x)}$$

The steps of algorithm are as follows :

1. Obtain all the Hindi translation equivalents for the first two English terms of the query.
2. Find the mutual information for the meaning of first two terms. Select the translated terms which have the highest mutual information for the first and the second query term respectively. Mark the selected translated term of second English query term as previous.
3. Obtain all the Hindi translation equivalents for the next English term of the query. Find the mutual information between previous and all translation equivalents of this term. Select the translated term which has the highest mutual information. Mark the selected translated term as previous.
4. Repeat step 3 until the meaning for all the terms not been found.

An automated process is developed for collecting the co-occurrence frequency from the test collection. The window size is the size of the document. The occurrence of each word in each document is counted, and then mutual information is applied for calculating the co-occurrence factor. We have used following strategies in English-Hindi CLIR experiments

**Strategy I (COOCC-I) :** Query submitted directly without using NLP tools

**Strategy II (COOCC-II) :** Transliterator is used for OOV.

**Strategy III (COOCC-III):** Stemmer and Transliterator are used but stopword are not removed from English query.

**Strategy IV (COOCC-IV) :** Stemmer and Transliterator are used Stop words are also removed from the English query.

The mean average precision of above four strategies were found to be 0.4407, 0.4672, 0.4580 and 0.4787 respectively. Comparing the performance of Strategy I and II, we find that the performance is higher when OOV words are transliterated. We further observe that the performance was best in Strategy IV in which stop words are removed, stemmer and transliterator are also used. A comparative recall precision graph is shown in Figure 3. A sample formulated query in Hindi is shown in the Table 2.

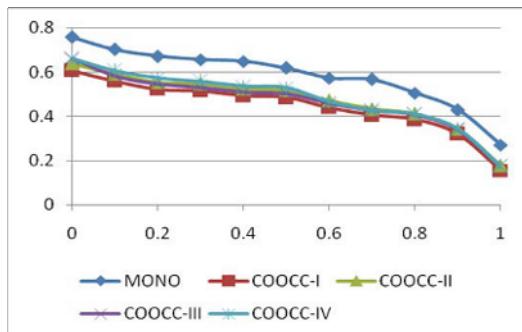


Fig. 3. Comparative recall and precision graph

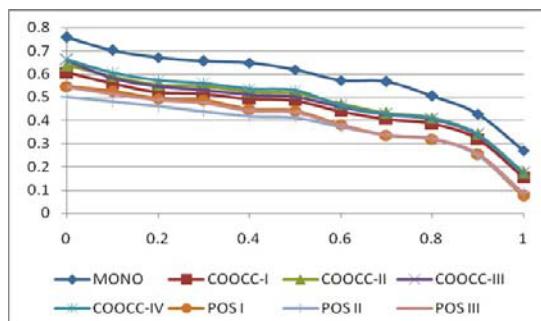
## 7 Conclusion

Dictionary-based method for CLIR is attractive because of cost effectiveness and easy to perform. The performance of dictionary based approach may be below than the monolingual retrieval as there is no one to one mapping found from source language to target language. The translation ambiguity and target polysemy

are the two major problems in dictionary based CLIR. The authors have evaluated part of speech based disambiguation strategy for overcoming the problem of target polysemy. The query formulation tried to deal with the problem of target polysemy, by finding the part of speech of a word in given context and then translation was performed. The query-by-query analysis shows that some of the queries were not translated properly in all the POS runs. The statistical technique is used for overcoming the problem of translation ambiguity in English-Hindi CLIR system. The results obtained in COOCC-IV run were best in the experiments performed. The comparative recall precision graph for all POS-runs and for all COOCC-runs is shown in Figure 4. It is worth mentioning that co-occurrence based technique performed better than the part of speech based query disambiguation. The query formulation was nearly perfect in COOCC-IV run and the mean average precision was 0.4787 and performance was 84% of the monolingual.

**Table 2.** Sample result of automatic translation result

English Query	Translated Query in COOCC-I	Translated Query in COOCC-II	Translated Query in COOCC-III	Translated Query in COOCC-IV
Right to Information Act	सही तक सूचना अधिनियम	सही तक सूचना अधिनियम	सही तक सूचना अधिनियम	अधिकार सूचना अधिनियम



**Fig. 4.** Comparative recall and precision graph

## References

1. Douglas W. : A Comparative Study of Query and Document Translation for Cross-Language Information Retrieval, Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup, pp. 472–483 (1998)
2. Hsin-Hsi Chen, Guo-Wei Bian and Wen-Cheng Lin,: Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval in proceedings of 27th Annual Meeting of the Association for Computational Linguistics, Univeristy of Maryland, College Park, Maryland, USA, ACL (1999)
3. Ballesteros L., Croft B. : Dictionary Methods for Cross-Lingual Information Retrieval. 7th DEXA Conf. on Database and Expert Systems Applications. Pages 791–801(1996)
4. Ballesteros L. , Bruce C.W.: Resolving Ambiguity for Cross-language Retrieval. In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (1998)
5. Pirkola A. :The effects of query structure and dictionary setups in dictionary-based cross-language information retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 55–63 (1998)
6. Davis M., Dunning T.: Query Translation using Evolutionary Programming for Multilingual Information Retrieval. The 41h Evolutionary Programming Conf., (1995).
7. Hull. D.A.: Using structured queries for disambiguation in cross-language information retrieval. In Proc. of AAAI spring symposium on cross-language text and speech retrieval, Stanford, CA (1997)
8. Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou, Changning Huang : Improving Query Translation for Cross-Language Information Retrieval using Statistical Models In Proceeding of 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (2001)
9. Sadat F., Maeda A., Yoshikawa M., Uemura S. : A Combined Statistical Query Term Disambiguation in Cross-Language Information Retrieval, Proceedings of the 13th International Workshop on Database and Expert Systems Applications (DEXA'02) 1529–4188/02 (2002)
10. Clough Paul, and Mark Stevenson,: ``Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Information Retrieval" In: Proceedings of the Second Global WordNet Conference , pp. 97–105 (2004)
11. Adriani M., van Rijsbergen C.J., : Term Similarity Based Query Expansion for Cross Language Information Retrieval. In Proceedings of Research and Advanced Technology for Digital Libraries, Third European Conference (ECDL'99), p. 311–322. Springer Verlag, Paris, September (1999)
12. Kekäläinen J., Järvelin K. : The impact of query structure and query expansion on retrieval performance. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia (1998)

13. Davis M.W., Ogden W.C. : Free Resources And Advanced Alignment For Cross-Language Text Retrieval. TREC 1997: 385–395(1997)
14. Monz C., Dorr B.J. : Iterative translation disambiguation for cross-language information retrievalin Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (2005)
15. Seetha A., Das S., Kumar M. : Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method. In proceedings of 10th International Conference on Information Technology (ICIT 2007), <http://doi.ieeecomputersociety.org/10.1109/ICIT.2007.40>
16. Daqing He, Oard D.W., Wang J., Jun Luo, Demner-Fushman D., Darwish K., Resnik P., Khudanpur S., Nossal M., Subotin M., Leuski A. : Making MIRACLEs: Interactive translingual search for Cebuano and Hindi September ACM Transactions on Asian Language Information Processing (TALIP), Volume 2 Issue 3 (2003)
17. Pingali P., Varma V. : IIIT Hyderabad at CLEF 2007 – Adhoc Indian Language CLIR task 2007 CLEF-2007, Cross Language Evaluation Forum 2007 Workshop at Budapest Hungary, At Eleventh European Conference on Digital Libraries (2007).
18. Mandal D., Dandapat S., Gupta M., Banerjee P., Sarkar S.: Bengali and Hindi to English Cross-language Text Retrieval under Limited Resources in CLEF 2007 working notes (2007).
19. Davis M.W., Ogden W.C. : Free Resources And Advanced Alignment For Cross-Language Text Retrieval.TREC: Gaithersburg, Maryland, 385–395(1997)
20. Seetha A., Das S., Kumar M., : Construction of Hindi test collection for CLIR research. In Proceedings of International Conference on Cognitive Systems (ICCS 2004) New Delhi, December 14–15, (available at [www.niitcrcs.com/iccs/iccs2004/Papers/240%20Anurag%20Sheetha.pdf](http://www.niitcrcs.com/iccs/iccs2004/Papers/240%20Anurag%20Sheetha.pdf)) (2004)
21. Croft W.B., Cook R., Wilder D : Providing Government Information on the Internet: Experiences with THOMAS. in Proceedings of DL. pp. 19–24 (1995)
22. Kamps J, Monz C., Maarten de Rijke, Sigurbjörnsson B. : Monolingual Document Retrieval: English versus other European Languages. In Proceedings of the Fourth Dutch Belgian Information Retrieval Workshop (DIR-2003). Pages: 35–39 (2003)
23. Porter M.F. : An algorithm for suffix stripping, in Program – automated library and information systems, 14(3): 130–137(1980)
24. [www.iit.net/trc/Dictionaries/Dict\\_Frame.html](http://www.iit.net/trc/Dictionaries/Dict_Frame.html)
25. Demner-Fushman D., Oard D. W. : The effect of bilingual term list size on dictionarybased cross-language information retrieval. In 36th Annual Hawaii International Conference on System Sciences (HICSS'03) – Track 4. Hawaii (2003)
26. Larkey L. S., Allan J., Connell, M. E., Bolivar A., Wade, C. : UMass at TREC 2002: Cross language and novelty tracks The 11th Text Retrieval Conference TREC 2002 NIST (2003)
27. ITRANS Indian language transliteration package at [www.aczone.com/itans](http://www.aczone.com/itans).
28. <http://nlp.stanford.edu/software/tagger.shtml>

# Evaluating Effect of Stemming and Stop-word Removal on Hindi Text Retrieval

Amaresh Kumar Pandey<sup>1</sup> and Tanvver J Siddiqui<sup>2</sup>

<sup>1</sup> Hughes Systique Corporation, Sec-33, Infocity,gurgaon, India  
amaresh.pandey@hsc.com

<sup>2</sup>Indian Institute of Information Technology, Deoghat,  
Allahabad, Uttar Pradesh, India-211012  
tanveer@iiita.ac.in

**Abstract.** IR system mainly use stop word elimination and stemming in indexing. This paper investigates the impact of stop word removal and stemming on Hindi Information Retrieval (IR). Three different stemmers have been used in this study and their performance has been compared. The experiments have been conducted on a test collection constructed using Hindi documents from EMILLE corpus. We created a stop-word list of Hindi by extracting the high frequency words from the collection and some manual addition. The evaluation has been made in terms of precision, recall and reduction of index size. The experimental investigation suggests that stop word removal improves retrieval significantly. However, we experienced a small drop in retrieval precision with all the three stemmer.

## 1 Introduction

Recent years have experienced a rapid growth of text material in Asian languages. Because English has been the main language for IR system development, much research focuses on it. Very little amount of work involves Asian languages. Although special tracks have been organized for Chinese in the TREC conferences, the results cannot be applied on other Asian languages except a few as they differ a lot linguistically. Particularly, Indian languages are known poorly from IR perspective. Among the Indian languages, Hindi and Bengali are listed in the top 10 most spoken languages of the world according to *Ethnologies list of most spoken languages*. Because of fast propagation of the Internet in South Asia over the last decade, the digital documents in regional languages are now available considerably. There are more than 60 online daily news publications, and other online Indian Language data sources like blogs, magazines, etc. on the Web [17]. Despite of this fact, not much has been done for supporting these languages during information retrieval over the Internet. However, in recent years many Indian language search engines are available on Internet. Also, the Government of India

has launched a mega digital library initiative and a country-wide cross-lingual information access consortium has also been established. The need for effective information access methods for Indian languages is therefore the need of present scenario. Most of the IR techniques that have proven successful for English may be ported to Indian Languages in a straightforward way, still more experimentations are required to make them fool proof. Although Indian language information retrieval (ILIR) research is in a novice state especially with regard to large-scale quantitative evaluation, several research efforts in this area have been reported in the recent past.

Publicly available tools for Hindi, one of the largest spoken languages in Asia, are scarce. IR systems for languages like Hindi, which differ from English in morphological behavior or in other features, cannot be developed properly without understanding the effect of stop word elimination and stemming. In this paper, we report our findings on the effect of stop word elimination and stemming on Hindi IR.

Stop word elimination and stemming are commonly used method in indexing. Stop words are high frequency words that have little semantic weight and are thus unlikely to help the retrieval process. Usual practice in IR is to drop them from index.

Stemming conflates morphological variants of words in its root or stem. It frees user from worrying about the truncation and inflection while framing queries and helps in reducing index size. Stemming does help in improving the retrieval performance. Particularly, recall is expected to improve after stemming.

When using stemming as a means to improve retrieval effectiveness one should be careful about under stemming and over stemming in choosing stemmer. Under stemming occurs when related words are not reduced to same stem. This may result in missing relevant document. Over stemming occurs when unrelated words are reduced to same stem thereby causing a match between query and irrelevant documents. A good stemmer has to find a right balance of conflation. A number of studies have been done on the impact of stemming on English IR and a number of stemmers are freely available. Unlike English, stop word lists and stemmers are not readily available for Hindi IR task. We used a list of 350 stop words in this work. We investigate the impact of stemming on Hindi IR with three different stemmers, namely light weight stemmer [9], UMass stemmer [12] and an unsupervised stemmer developed by us.

Due to limited resources for Hindi, we perform our study on a test collection developed using documents from EMILEE corpus [26]. Total number of documents in the test collection is 700 and Total number of queries used is 70. The average length of document in this test collection is larger than early English IR test collections. With small documents chances of mismatch is high, if no stemming is used.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 presents our approach for developing Hindi stemmer, stop-word list and retrieval strategies used for Hindi IR. Section 4 shows our experiments, data set generation for Hindi IR and analysis of results. We will conclude our paper in section 5.

## 2 Related Work

A number of studies have been conducted on the impact of stemming on IR. The results have been mixed. Harman [17] failed to find any improvement with simple stemmers for English language whereas Frakes and Hull [18, 19] reported a small benefit. Hull [19] compared a wide range of stemmers. The results may not be the same for other languages. Braschler and Ripplinger [20] found stemming and decompounding beneficial for German text retrieval. Other studies on stemming behaviors of languages other than English include Slovene [21], French [22], Italy [23], and German, Dutch and Italian [24]. The development and use of stemming algorithms are not new initiatives. First ever published stemmer was written by Julie Beth Lovins [10]. Another widely known stemmer was developed by Martin Porter [11]. These early stemmers were rule-based. Formulating such rules for a new language is time consuming. The Word Frame model proposed by Wicentowski [16] is a supervised approach for stemming. It requires a set of inflection-root pairs to learn a set of string transduction. These transductions are then used to reduce unseen inflections into their root. A lot of work on morphology focuses on unsupervised approaches [2, 13, 4, 5, 14]. [14 and 13] later proposed a Bayesian model for MDL for English and French. Goldsmith's [4] work is based on minimum description length principle which prefers a model that gives rise to minimal length coding of observed data set. Yet another work on stemming was reported by Freitag [15]. His work is based on automatic clustering of words using co-occurrence information. The suffixes have been identified using orthographic dissimilarity between the words in different clusters. Studies on stemming behavior for Indian languages suffer from scarce availability of suitable test collection. For English, well-known TREC collections exist. Comparable test collections become available only recently for major European and Asian languages through CLEF and NTCIR campaigns. Previous works on Hindi stemming behavior and development of stemmers for Indian languages include [12, 25, 6, 9]. Ramanathan and Rao [9] used a light weight stemmer consisting of 27 common suffixes and the UMAss stemmer developed by Larkey et al. [12] in their work. Both the stemmers contain suffixes extracted manually. They reported slight improvement in performance. The work performed by Chen and Gey [25] uses a statistical Hindi stemmer. They found no improvement for monolingual retrieval. Dasgupta and Ng [6] present an unsupervised morphological analyzer for Bengali. They reported a maximum accuracy of 64.6% when tested on 5000 words. A Telugu morphological generator, TelMore, is outlined by Ganapathiraju et al. [3]. TelMore is a rule-based approach. It uses the results of analysis of Telugu performed by [7, 8].

## 3 Our Approach

### 3.1 Stemmer

In this section, we have discussed three stemmers which we have considered for evaluation. First we have briefly explained our stemmer, than light weight stemmer and finally we have discussed the UMass stemmer.

Light weight stemmer is developed by Ramanathan, A. and Rao, D in 2003. It is based on observation of Hindi morphology of various categories like, noun inflections, adjective inflections and verbs inflections. Their list contains 65 suffixes. The reported accuracy figure in terms of over stemming and under stemming is 13.84% and 4.68% respectively. Table 2 shows the list of suffixes which is used in light weight stemming.

**Table 2.** Suffix list of light weight stemmer

न लि ली दु दू ले लो लें औं लां दुआं दुएं दुओं लाएं लाओं यिआं यिओं लयिआं लयिओं लाओं लीयाौं लैयाँ तरै नारै नाओं ता ती लीं तीं ते लाता लाती लातीं लाते ना नी ने लाना लाने ऊंगा ऊंगी लाऊंगा लाऊंगी एंगे एंगी लाएंगे लाएंगी लोगे लोगी लाओगे लाओगी लेगा लेगी लाएगा लाएगी लाया लाए लाई लाई लिए लाओ लाइए कार काकार
--

UMass stemmer developed by Larkey et all [12]. They performed experiments on cross-lingual information retrieval for Hindi-English language pair. Their suffix list consists of 27 suffixes for Hindi. Table 3 shows the list of suffixes which is used in UMass stemming.

**Table 3.** Suffix list of UMass stemmer

दूँ दूँ लैं लैं लों लों यों यों लिय ता ती ते ना नी ने के ला ले ली लो लेंगे लूँगा लूँगी लेगा लेगी लियाँ याँ
---

Details of the unsupervised stemmer can be find in [1]

### 3.2 Stop-Word -list Generation

We have used a list of 350 stop words. For generation of stop word list, we extracted 400 high frequency words from 1000 Hindi documents of EMILLE corpus. We dropped 90 words from this list and added 40 more words which seems good candidate for being called stop words but were not among 400 high frequency words in the corpus e.g. तुम्हारी, रहा, रही, है, थे, उसका, उसकी, भि, भी, अरे, ओह, न, इसे, वो, औ, तो, फिर, क्यूंकि... etc.

### 3.3 Document and Query Collection

Standard IR test collections for Hindi do not exist till date. For evaluation, we created an IR test collection consisting of 700 Hindi document randomly extracted from EMILEE corpus[26] and 70 queries. Documents in this collection are news articles from the India Info, Ranchi Express and Web-Dunia websites. Some documents have been taken from CIIL Corpus. Total number of words in the collection is approximately 12,390,000 words. All data are encoded in two-byte Unicode text.

In order to generate queries, five subjects were given 140 documents each. The subjects were M.Tech. second semester students. They were asked to read the document and provide 15 queries each along with the target documents. We manually checked the submitted queries for redundancy. Five queries found redundant were dropped. So, we were left with a total of 70 queries.

### 3.4 Hindi IR

The other issues involved in Hindi IR are:

*Document structure:* We have considered only <text> and <body> section of document for indexing purpose.

*Encoding scheme:* UTF-8 encoding scheme is used in this work.

*Noise removal:* Noise removal transforms document by removing header, footer and other text which are not related to subject of documents. The transformed document only contains the content of the document.

*Tokenization:* Tokenization of text implies breaking the text into separate lexical. This is the very first step of text processing. Word can be separated by white space, comma, dash, dot etc. A number of issue comes while tokenization. We have explained most of the issues in actual implementation section.

*Stop word list removal:* Stop word removal involves removal of all those words which do not contribute to meaning and which occur with high frequency and

most common words in the text. We maintain a list of stop word and use it while processing the text. By observation we found that around 30% of size will reduce by removal of stop words.

*Stemming:* Stemming is the process of removing inflection of words, so that we have to convert different words in the same root or stem word which contains the similar meaning, for that, we have developed our own morphological analyzer which is based on the unsupervised approach for Hindi.

*Term-weighting scheme:* Term weighting scheme is the fundamental issue in information retrieval task, earlier we have explained the different type of term weighting scheme. We have used the tf\*idf as term weighting scheme.

*Indexing strategies:* Indexing scheme is also one of the major fundamental issues in information retrieval task, earlier we have explained the different type of indexing scheme. We have used word as indexing scheme.

*Retrieval and ranking:* Retrieval of documents is based on cosine similarity between the query and document. Ranking of retrieved document is done by cosine values, document which gives the more cosine value will be ranked first.

*Evaluation:* We have done experimentation in terms of precision, recall, mean average precision and mean average interpolated precision.

### 3.5 Retrieval Strategies

Vector space retrieval model has been used in evaluation. The inner product between query and document vector is used to give a term similarity score. The retrieval strategy can be formally defined by the following expressions:

$$\text{Document-term weight: } W_{ij} = \frac{Tf_i}{\max}$$

$$\text{Query-term weight: } W_{ik} = t$$

$$\text{Query-document similarity: } \bullet w_{ij} \times w_{ik}$$

Where  $w_{ij}$  = weight of  $i^{\text{th}}$  term in  $j^{\text{th}}$  document.

$Tf_{ij}$  = frequency of  $i^{\text{th}}$  term in  $j^{\text{th}}$  document

## 4 Experimentation and Results

To see the impact of stop word elimination and stemming on Hindi IR, we performed test runs on a Hindi IR test collection. The baseline performance corresponds to no stemming with stop words removed. The details on the development of this collection are discussed in section 4.1.

#### 4.1 Test Collection

The IR test collection used in this work has been created using Hindi documents from EMILLE corpus [26] (section 3.2). Table 4 lists the characteristics of our test collection and table 5 gives the statistics of query collection.

**Table 4.** Statistics of documents collection

Number of Doc.	700
Avg. length of Doc.(no of term)	7801
Min length of Doc	271
Max length of Doc.	21612
Size of Doc.(total)	64 MB
Avg. size of Doc	94 KB

**Table 5.** Statistics of query collection

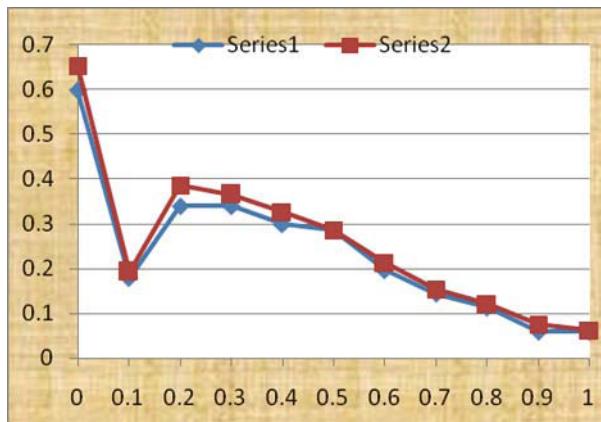
Total no. of query	70
Avg. no of relevant documents per query	10
Min number of relevant document	4
Max number of relevant document	45
Total coverage of document	90%

#### 4.2 The Experiment

The first test run investigates the impact of stop word elimination on retrieval. The evaluation measures used are mean average precision and mean average interpolated precision. Table 6 shows the result of this experiment. Fig. 1 shows the recall-precision curve for this experiment. Table 7 shows the reduction in index size obtained after stop word removal. The second experiment investigates the impact of stemming on Hindi retrieval. The details of the three stemmers used in this experiment have been provided in section 2. The results are shown in table 8. Fig 2 shows the recall-precision curve using 11-point interpolated mean average precision. The third experiment was performed to see the impact of stemming on index size. The % reduction obtained with three stemmers has been compared in table 9.

**Table 6.** Retrieval performance with and without stop-word removal

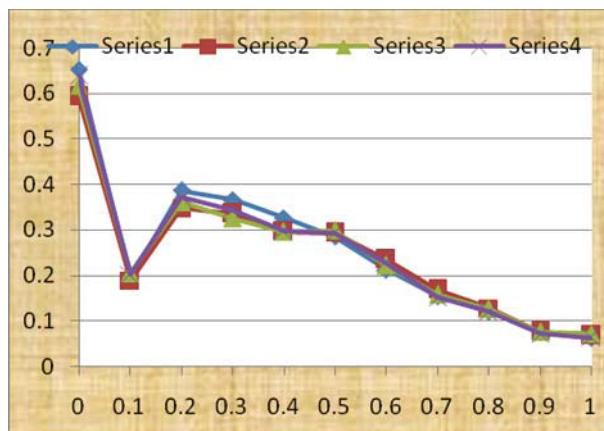
Stop-word	Mean Avg. Precision (MAP)	Mean Avg. Interpolated precision (MAIP)
With out stop-word removal	0.4083 0.4504	0.23775
With stop-word removal		0.257972

**Fig.1** 11-point interpolated precision-recall curve (1-with out stop-word removal, 2-stop word removal)**Table 7.** Reduction in index size

Stop-word	% reduction in number of unit	% reduction in size
With out stop-word removal	Reference	
With stop-word removal	22.45%	21.23%

**Table 8.** Experimental results with stemming in terms of MAP and MAIP

Stemmer	Mean Precision (MAP)	Mean Interpolated precision (MAIP)
With out stemming	0.4504	0.257972
Our stemmer	0.3964	0.248169
Light weight stemmer	0.4025	0.250625
UMass Stemmer	0.4291	0.252722

**Fig. 2.** Recall-precision curve (1-nostemming, 2-unsupervised stemmer, 3-Light weight stemmer, 4- UMass stemmer)**Table 9.** Experimental results with stemming in terms of size reduction

Stemmer	% reduction in number of unit	% reduction in size
Our stemmer	<b>18.24 %</b>	<b>15.01%</b>
Light weight stemmer	12.348%	9.12%
UMass Stemmer	15.25%	12.41%

### 4.3 Results and Discussion

As shown in Table 8, all the three stemmer degrade retrieval performance. The mean average precision without stemming was found to be 0.45 whereas with unsupervised stemmer, light weight stemmer and UMass stemmer it was found 0.39, 0.40 and 0.43 respectively. The best performance was observed with UMass stemmer with a decrease of 4.4 % as compared to baseline performance. The drop with unsupervised and light weight stemmer was 13.33 % and 11.11 % respectively. We feel that this drop is due to over stemming in certain cases.

As shown in Table 6, the observed mean average precision with and without stop-word elimination is .4083 and 0.4504 respectively and in Table-7, we observed after removal of stop-word percentage reduction of index size is 22.45%. Experimental investigation show stop-word removal give better results in terms index size reduction as well as retrieval effectiveness.

## 5 Conclusion

This paper investigated the impact of stop word elimination and stemming on Hindi IR. The empirical investigations suggest that stop word elimination improves retrieval performance significantly. However, all the three stemmer degrade the performance. The UMass stemmer performed well as compared to our stemmer and light weight stemmer. The performance of light weight stemmer and the unsupervised stemmer developed by us was quite close with a % and % drop respectively in mean average precision. Due to limited resources for Hindi, we perform our study on a test collection on a small test collection. Large IR test collection is needed to draw more reliable conclusion. Further, we used a simple vector space model in this study; other indexing strategies need to be investigated.

## References

1. Pandey, A., Siddique, T.: "An unsupervised Hindi stemmer with heuristic improvements", Proceedings of AND 08, Singapore, pp 99–105.
2. Bharati, Sangal A.R., Bendre S.M., Kumar P., Aishwarya : Unsupervised Improvement of Morphological Analyzer for Inflectionally Rich Languages, Proceedings of the NLPRS, pp 685–692 (2001)
3. Ganapathiraju, Madhavi A, Levin L. : TelMore: Morphological Generator for Telugu Nouns and verbs, In the proceedings of Second International Conference on Universal Digital Library Alexandria, Egypt November 17–19 (2006)
4. Goldsmith, J.. : Unsupervised Learning of the Morphology of a Natural Language. Computational Linguistics, 27, 153–198, (2001)
5. Creutz, M., Lagus, K. : Unsupervised Morpheme Segmentation and Morphology Induction from Text Corpora Using Morfessor 1.0.Tech. rep. A81, Helsinki University of Technology, (2005).

6. Dasgupta S., Vincent N.: Unsupervised morphological parsing of Bengali, Brown, C.P., In: The Grammar of the Telugu Language.1991, New Delhi: Laurier Books Ltd. (2007)
7. Krishnamurti, B.; A grammar of modern Telugu. Delhi ; New York: Oxford University Press (1985).
8. Ramanathan, A. Rao, D. : A lightweight stemmer for Hindi. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL), on Computational Linguistics for South Asian Languages (Budapest, Apr.) Workshop. (2003)
9. Lovins J.B. : Development of a stemming algorithm. Mechanical Translation and Computational Linguistics 11:22–31. (1977)
10. Porter, M.; An algorithm for suffix stripping program. Vol. 14, pp. 130–137(1980)
11. Larkey L.S., Connell M.E., Abduljaleel N. : Hindi CLIR in Thirty Days. ACM Transaction on Asian Language Information Processing, Vol-2, No. 2, , Pages No. 130–142 (June 2003)
12. Snover, M.G., Brent, M. R. : A Bayesian model for morpheme and paradigm identification. In Proceedings of the 39th annual meeting of the ACL, pp. 482–490. (2001)
13. Brent, M. R., Murthy, S. K., Lundberg, A. : Discovering morphemic suffixes: A case study in minimum description length induction. In Proceedings of the fifth international workshop on artificial intelligence and statistics (1995)
14. Freitag, D. : Morphology induction from term clusters. In Proceedings of the ninth conference on computational natural language learning (CoNLL), pp. 128–135. (2005)
15. Wicentowski R. : Multilingual Noise-Robust Supervised Morphological Analysis using the WordFrame Model. In Proceedings of Seventh Meeting of the ACL Special Interest Group on Computational Phonology (SIGPHON), pp. 70–77, (2004)
16. Harman, D. : How effective is suffixing? Journal of the American Society for Information Science, 42(1),7–15. (1991)
17. Frakes, W.B. : Stemming algorithms. In: Frakes, W.B. and Baeze-Yates, R. (editors) Information Retrieval: Data Structures and Algorithms. Englewood Cliffs: Prentice-Hall, pp. 131–160 (1992)
18. Hull D.A. : Stemming algorithms: a case study for detailed evaluation. Journal of the American Society for Information Science, v.47 n.1, p.70–84, Jan. (1996)
19. Braschler M., Ripplinger B. : How Effective is Stemming and Decompounding for German Text Retrieval? Inf. Retr. 7(3–4): 291–316 (2004)
20. Popovic M., Willett P. : The Effectiveness of Stemming for Natural-Language Access to Slovenc Textual Data. JASIS 43(5): 384–390 (1992)
21. Savoy J: A Stemming Procedure and Stopword List for General French Corpora. JASIS 50(10): 944–952 (1999)
22. Sheridan P., Ballerini J.P. : Experiments in Multilingual Information Retrieval Using the SPIDER System. SIGIR: 58–65 (1996)
23. Kamps J., Monz C., Maarten de Rijke :Combining Morphological and Ngram Evidence for Monolingual Document Retrieval In: M.-F. Moens, R. De Busser, D. Hiemstra, W. Kraaij, editors, Proceedings of the Third Dutch Information Retrieval Workshop (DIR 2002), pages 47–51
24. Chen, A. and Gey, F.C. ;: Generating statistical Hindi stemmers from parallel texts. ACM Trans. Asian Language Inform. Process. Vol. 2, No. 3, Sep. (2003)
25. The EMILLE Corpus, <http://bowland-files.lancs.ac.uk/corplang/emille/>

# An Unsupervised Approach to Hindi Word Sense Disambiguation

Neetu Mishra, Shashi Yadav and Tanveer J. Siddiqui

<sup>1</sup>Indian Institute of Information Technology, Allahabad. UP, India.  
{neetumishra, tanveer,}@iiita.ac.in

**Abstract:** This paper presents an unsupervised word sense disambiguation algorithm for Hindi. The algorithm learns a decision list using untagged instances. Some seed instances are provided manually. Stemming has been applied and stop words have been removed from the context. The list is then used for annotating an ambiguous word with its correct sense in a given context. The evaluation has been made on 20 ambiguous words with multiple senses as defined in Hindi WordNet. Total training instances are 1856 and total test instances are 1641. The performance has been measured in terms of accuracy. The experimental investigation suggests that stop word removal and stemming improves performance of the algorithm.

## 1 Introduction

Natural languages contain words that have multiple senses or meaning. Human beings easily recognize the correct meaning of a word without even considering all of its senses. However, it creates problem during automatic processing of text. In order to get the correct meaning of a word disambiguation has to be performed. The task of disambiguation is concerned with identifying correct meaning of an ambiguous word in a specific use. For example, consider the word “शाखा” in the following sentence.

प्रकृति के इस अनुरागभरे आङ्गन को हम समझ सकें तो हमें प्रत्येक वृक्ष, शाखा, फूल और पत्ते-पत्ते से अपने नाम की पुकार सुनाई देगी।

In this sentence the word शाखा is ambiguous; it has 4 senses as listed in Hindi WordNet<sup>1</sup> (Fig. 1). A simple dictionary lookup operation will not get the intended meaning.

---

<sup>1</sup> <http://www.cfilt.iitb.ac.in>

1. (R) शाखा, डाल, डाली, शाख, शाख, साख, साखा - वृक्ष आदि के तने से इधर-उधर निकले हुए अंग "बच्चे आम की शाखाओं पर झूल रहे हैं"
2. (R) सम्प्रदाय, संप्रदाय, पंथ, मत, पाषड़, पाषण्ड, शाखा - कोई विशेष धार्मिक मत या प्रणाली "वह शैव सम्प्रदाय का अनुयायी है", किसी विषय या सिद्धांत के संबंध में एक ही विचार या मत रखनेवाले लोगों का वर्ग "जैन धर्म के अंतर्गत दो शाखाएँ हैं-दिगंबर और शेतांबर"
3. (R) अंग, शाखा, घटक, अययव, संघटक - किसी वर्ग विशेष का घटक या भाग जो अपने आप में पूर्ण भी होता है "इस संस्था के पाँच अंग हैं"
4. (R) शाखा - किसी बड़े या अत्यधिक जटिल संस्था का विभाग "इस दवा कंपनी की कई शाखाएँ हैं"

**Fig. 1.** Senses of शाखा obtained from the Hindi WordNet

Most of the work on word sense disambiguation (WSD) focuses on English. Very little amount of research has been done on automatic word sense disambiguation for Hindi [2]. The work in this A classifier is learned from unlabelled Hindi corpus. The approach is inspired from Yarowsky's work[5] but differs from it in the following points:

- (i) Unlike the work in [2], we create seed instances automatically.
- (ii) We remove stop words from the context and then perform stemming on remaining words. This takes care of the case when the word appearing in the context is morphological variant of a word in the decision list.
- (iii) Unlike Yarowsky's work, this work considers more than two senses.

The performance of supervised approaches has been found better than unsupervised approaches [5], however, they require sense tagged corpora for training. Building such a corpus is a time consuming task. To the best of our knowledge no known sense tagged corpus is available for Hindi. Hence this focuses on unsupervised approach. A small test corpus consisting of 20 ambiguous words has been developed for evaluation. Various senses of these words have been defined using Hindi WordNet.

The structure of the paper is as follows:

Section 2 gives insight to various approaches for WSD. It explains in detail different approaches used to solve WSD mainly supervised and unsupervised the work related to unsupervised WSD in English and as well as Hindi. Section 3 and 4 describe about Yarowsky's algorithm and some details about our

implementation. It also includes a brief introduction about the tool Hindi WorldNet used in Hindi WSD. Proposed Approach explains in detail the approach adopted for WSD. It also explains the training/test datasets used by the Hindi Word Sense Disambiguation System. Section 5 presents the detail of the experiments conducted. Finally, we conclude in section 6.

## 2 Related Work

The automatic disambiguation of word senses has been a matter of interest since the earliest days of computer understanding of natural language in the 1950's. Sense disambiguation is an "intermediate task" ([3] for a number of natural language processing applications such as content understanding, machine-man communication, machine translation, etc. WSD approaches can be broadly classified as dictionary-based and corpus-based. Dictionary-based approaches make use of lexical resources, e.g. dictionary, thesaurus, ontology, etc., for disambiguation whereas corpus-based approaches make use of large corpus to gather sense information. An overview of early WSD approaches can be found in [1].

The work in [6] is in line with dictionary-based approach in which correct sense of a word is identified by matching the context of ambiguous words with its dictionary definition. The sense having maximum overlap is assigned to target word.

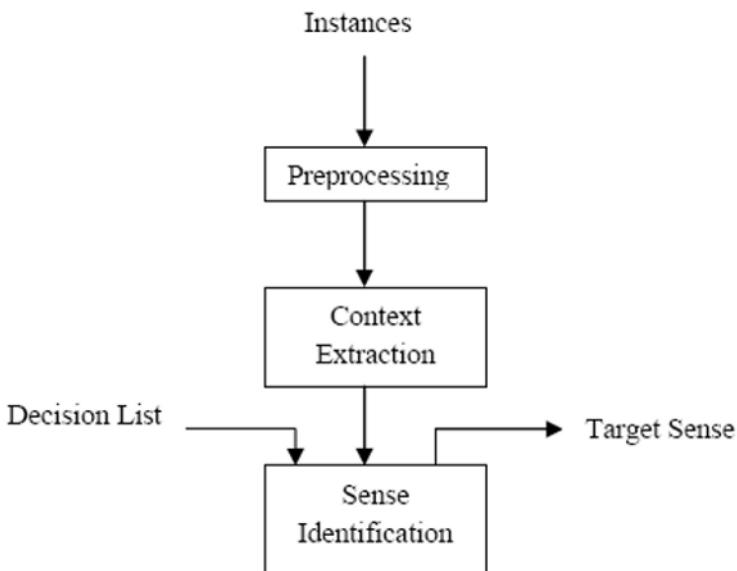
Yarowsky extended the idea by combining evidences from thesaurus and supervised learning [4]. Other works in supervised disambiguation include [8] [12]. Creating sense tagged corpus required by supervised approaches is quite time consuming. Yarowsky [5] proposed an unsupervised approach for disambiguation which uses unlabelled text for training. It can be easily extended for languages for which sense tagged corpus is not available. Unsupervised algorithm broadly fall in two categories: similarity based & graph based. Similarity based algorithm utilize surrounding context to disambiguate a word whereas graph based algorithm work by building a graph and identifying the most important node for each word. Nodes in the graph correspond to word senses and edges correspond to dependencies between them. A comparative study of these two types of algorithm has been made in [9] and [10].

Bernard and Johnson [13] introduced a model for word sense disambiguation that uses images for disambiguating a word. They evaluated their approach using a corpus containing coral image dataset associated with disambiguated text from Semcor corpus and suggested that visual information can help in WSD. Lin and Versoor [11] proposed semantics enhanced language model for unsupervised WSD. The language model used in this work extends n-gram model with semantics.

Sinha et al. [2] used Hindi WordNet<sup>2</sup> for sense identification. This paper presents an unsupervised approach for Hindi word sense disambiguatuion.

### 3 Our Approach

The architecture used in unsupervised approach has been shown in Figure. 2. First, we pre-process the raw text to remove stop words and to reduce morphological variations of words to their stem. Then, for each sense collocations appearing in its sense definitions in Hindi Word Net are identified. These seed instances have been shown in table1 are used to label unlabelled instances resulting in a growing number of sense tagged instances. These new tagged instances are further used to identify collocates that help in tagging some of the remaining instance. The collocations are ranked on the basis of probability value. The process is iterated until convergence is achieved. The resulting list is used for assigning tags to new contexts.



**Fig. 2.** System architecture

The input to be disambiguated is the text containing an ambiguous word and the target word. The ranked decision list learned from unlabeled text is applied on it to get the correct sense.

1. Pre-process the corpus to remove stop words and to reduce words to their stems.
2. Create training text by extracting context of all use of ambiguous words.
3. For each sense of a target word, extract collocations which are reliable indicator of that sense using Hindi Word Net.
4. Use collocates identified in step 2 to classify the context to obtain seed set.
5. Run supervised disambiguation algorithm on seed set to create a decision list classifier. The decision list consists of collocations ranked on the basis of their probability value.

$$P(\text{sense}_i / \text{collocate})$$

6. Apply the classifier on the remaining unlabeled text. The newly tagged instances are added to seed set and new indicators (collocations) are identified.
7. Repeat step 4-6 until no change occurs in the remaining unlabeled text.
8. The resulting classifier can now be applied on new data.

**Table 1.** Initial seed word

Sense id	Seed word for corresponding sense id
1	पौधों
2	संप्रदाय
3	विज्ञान

अंग ,कर ,कलम, कुंभ ,  
 खाता, गोला ,ढाल ,दर ,  
 फल ,पद ,बाली ,तीर ,  
 मत ,माँग ,शाखा ,स्तंभ,  
 सोना ,सीमा ,हल ,श्रेणी

**Fig. 3.** Ambiguous words

#### 4. Data-set Generation

We have developed our own training and test corpus which includes the Hindi corpus from IIT Bombay and news articles from the India Info, Ranchi Express, Dainik Jagran, Hindi Wikipedia and Web-Dunia websites. Some documents have been taken from EMILLE Corpus. All data are encoded in two-byte Unicode text. Figure. 3 shows list of ambiguous words considered in this work. Table 2 and Table 3 give the statistics of the collection.

The senses of target words have been defined using Hindi WordNet. The training document for each target word is selected. Average number of instances per word collected for training is 90 and average number of instances per word per sense is approximately 30. The test data set has an average number of 80 instances per word and average number of 25 instances per word per sense. Table 4 shows the details of words and their senses used in this paper.

**Table 2.** Training set

Total number of ambiguous words	20
Average size(number of instances) of the training data /word	90
Average size(number of instances) of the test data /word	80

**Table 3.** Test set

Total number of ambiguous words	20
Average size(number of instances) of the training data /word /sense	30
Average size(number of instances) of the test data /word /sense	25

**Table 4.** Words with number of senses

Number of senses	Target words
2	खाता , ढाल, पद, तीर, सोना, सीमा, हल
3	अंग, कर , कुभि , दर, फल, बाली, शाखा, श्रेणी
4	कलम , मत , मँग
5	गोला , स्तंभ

## 5. Experiment and Results

Two test runs have been made on 20 ambiguous Hindi words to assess the performance of the proposed approach. The accuracy has been measured in terms of correctly classified instances. The first test-run observes the results with and without stopword. The objective of the second test run is to assess the impact of stop word removal and stemming.

$$\text{Accuracy} : \frac{\text{Number of correctly classified contexts}}{\text{Total number of context considered}}$$

We have collected results on the basis of the way decision list has been created ie manuaaly and automatically and one after applying stemming. Both the test runs have been made for 20 target words. An average accuracy of 86.3%, 87.2%, 89.5% and 91.8% respectively in RUN-1, RUN-2 has been achieved. By using a stemmed decision list we can even classify the sense of target word whose context has any morphological variant of stemmed word in decision list. The Variation in performance is due to the morphological variation in corpus. A subset of the target words and their accuracies has been listed in Figure 4.

Word	Accuracy	Word	Accuracy
अंग	0.96	कर	0.95
कलम	0.98	कुभ	0.96
खाता	0.96	गोला	0.95
ढाल	0.98	दर	0.97
फल	0.94	पद	0.90
माँग	0.92	मत	0.92
शाखा	0.96	स्तंभ	0.96
सोना	0.94	सीमा	0.97
हल	0.98	श्रेणी	0.94

**Fig. 4.** Word and their accuracies

## 6 Conclusion and Future Work

We have achieved an average accuracy of 91.8% on 20 words which seems promising. As the approach does not require sense tagged data for training, it can be applied easily on other resource poor languages as well. However, we would like to test the proposed approach on more extensive data set before making any general conclusion. The accuracy of the system varies from about 82% to about 92% with stop word removal, automatic decision list generation and with stemming.

## References

1. Ide N., Véronis J. :Word sense disambiguation. The state of the art (1998)
2. Sinha M., Kumar M., Pande P., Kashyap L., Bhattacharyya P.;Hindi Word Sense Disambiguation. International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, (2004)
3. Stevenson, M., Yorick, W. : The Interaction of Knowledge Sources in Word Sense Disambiguation. Computational Linguistics, 27:3, 321–349 (2001)
4. Yarowsky, D. :Word-sense disambiguation using statistical models of Roget's categories trained on large corpora". In: Proceedings of the 14th International

- Conference on Computational Linguistics (COLING-92), pp. 454–460, Nantes, France (1992)
- 5. Yarowsky D. : “Unsupervised Word Sense Disambiguation Rivaling Supervised Methods”, Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, MA, pp. 189–196. (1995)
  - 6. Lesk, Michael : Automated Sense Disambiguation Using Machine-readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone.” Proceedings of the 1986 SIGDOC Conference, Toronto, Canada, June 1986, pp. 24–26. (1986).
  - 7. Fellbaum C., Alkhalifa M., Black W., Elkateb S., Pease A., Rodriguez H., Vossen P. : Domain-Specific Word Sense Disambiguation. In: Eneko Agirre and Philip Edmonds (eds) Word Sense Disambiguation - Algorithms and Applications. Springer, June (2006).
  - 8. W.Gale, K. Church, and D. Yarowsky. : One sense per discourse. In Proceedings of the DARPA Speech and Natural Language Workshop, pp. 233–237, Harriman, NY, February.( 1992)
  - 9. Mihalcea R. : Unsupervised large-vocabulary word sense disambiguation with graph based algorithms for sequence data labeling. In: Proceedings of HLT/EMNLP, pp. 411–418,Vancouver,BC (2005)
  - 10. Brody et al, Brody S., Navigli R., Lapata M. : Ensemble methods for unsupervised WSD.In Proceedings of the ACL/COLING, Sydney, Australia, (2006).
  - 11. Shou-de Lin, Verspoor K. : A Semantics-Enhanced Language Model for Unsupervised Word Sense disambiguation in LNCS Volume 4919, 287–298, Feb (2008).
  - 12. Mooney R. J.: Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning, In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP), pp.82–91, (1996).
  - 13. Barnard K., Johnson M. : Word sense disambiguation with pictures, Artificial Intelligence 167, 3–30 (2005)

# Search Result Clustering using a Singular Value Decomposition (SVD)

Hussam Dahwa Abdulla and Vaclav Snasel

*Dept. of Computer Science, VSB – Technical University of Ostrava  
hussam.dahwa@hotmail.com, vaclav.snael@vsb.cz*

**Abstract:** There are many search engines in the web, but they return a long list of search results, ranked by their relevancies to the given query. Web users have to go through the list and examine the titles and (short) snippets sequentially to identify their required results. In this paper we present usage of Singular Value Decomposition (SVD) as a very good solution for search results clustering.

## 1 Introduction

In the last few years the world observes exponential growing of the amount of information. Easiness of using this information and easiness of access to this information brew a big problem to retrieval of information, and the results contain a lot of data and it can be hard to choose or find the relevant information in the result. The huge numbers of data and inability to recognize the type of data lead to inability for the right searching for information. For users with no prior experience, searching for topic manually in the web can be difficult and taking time.

The major difficulties are the complicity of the content and the classification of the huge information in the web, and identifying and naming topics and relationships between these topics. In this situation, clustering data gives us a good result for data analysis. We can use the search result clustering in width area from different fields. In this paper we present one of the methods for clustering data to be used in the search result clustering. We use the singular value decomposition as a mathematical method to reduce a big value of objects by combining the attributes of these objects [1].

## 2 Search Results Clustering

In the last years, the search result clustering has attracted a substantial amount of research (e.g., information retrieval, machine learning, human-

computer interaction, computational linguistics, data mining, formal concept analysis, graph drawing).

Search result clustering groups search results by topic. Thus provides us with complementary view to the information returned by big documents ranking systems. This approach is especially useful when document ranking fails to give us a precise result. This method allows a direct access to a subtopic; search result clustering reduces the information, helps filtering out irrelevant items, and favours exploration of unknown or dynamic domains. Search result clustering, is different from the conventional document clustering. When clustering takes place as a post-processing step on the set of results retrieved by an information retrieval system on a query, it may be both more efficient, because the input consists of few hundred of snippets, and more effective, because query-specific text features are used. On the other hand, search result clustering must fulfil a number of more stringent requirements raised by the nature of the application in which it is embedded[2].

### 3 Singular Value Decomposition (SVD)

Singular Value Decomposition (SVD) breaks a  $n \times m$  matrix A into three matrices U,  $\Sigma$  and V such that  $A = U\Sigma V^T$ . U is a  $(n \times k)$  orthogonal matrix whose column vectors are called the left singular vectors of A, V is a  $(k \times m)$  orthogonal matrix whose column vectors are termed the right singular vectors of A, and  $\Sigma$  is a  $(k \times k)$  diagonal matrix having the singular values of A ordered decreasingly. Columns of U form an orthogonal basis for the column space of A.

Singular value decomposition (SVD) is well-known because of its application in information retrieval as LSI. SVD is especially suitable in its variant for sparse matrices. [7][8][9].

Since only the first k concepts can be considered are semantic important (the singular values are high), we can approximate the decomposition as

$$A = U_k \Sigma_k V_k^T$$

where  $U_k$  contains the first k most important concept vectors,  $\Sigma_k$  contains the respective singular values and  $\Sigma_k V_k^T$  contains the pseudo-document vectors represented using the first k concept vectors. In other words, by SVD the original m-dimensional vectors are projected into a vector space of dimension k ( $k \ll m$ ). The SVD approximation (so-called rank-k SVD) can be created either by “trimming” the full- SVD matrices or by usage of a special method designed to perform directly the rank-k SVD.

Theorem: (Singular Value Decomposition) Let A be an  $m \times n$  rank-r matrix. Be  $\sigma_1, \sigma_2, \dots, \sigma_r$  eigenvalues of a matrix  $\sqrt{AA^T}$ . There exist orthogonal matrices U ( $u_1$

, . . . ,  $u_r$ ) and  $V(v_1, \dots, v_r)$  for  $r =$ , whose column vectors are orthonormal, and diagonal matrix  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ .

The decomposition  $A = U \Sigma V^T$  is referred to as a singular value decomposition of matrix A and numbers  $\sigma_1, \dots, \sigma_r$  are singular values of the matrix A. Columns of U (or V) are referred to as left (or right) singular vectors of matrix A.

Now we have a decomposition of the original matrix A. Needless to say, the left and right singular vectors are not sparse. We have at most r nonzero singular numbers, where rank-r is the smaller of the two matrix dimensions. Because the singular values usually decrease quickly, we only need to take k greatest singular values and corresponding singular vector coordinates and create a k-reduced singular decomposition of matrix A.

**Definition:** Let us have  $k, 0 < k < r$  and singular value decomposition of A

$$A = U \Sigma V^T = (U_k U_0) \begin{pmatrix} \Sigma_k & 0 \\ 0 & \Sigma_0 \end{pmatrix} \begin{pmatrix} V_k^T \\ V_0^T \end{pmatrix}$$

$A = U_k \Sigma_k V_k^T$  is referred to as a k-reduced singular value decomposition (k-rank SVD).

In information retrieval, if every document is relevant to only one topic, we obtain a latent semantics – semantically related words and documents will have similar vectors in the reduced space. For an illustration of rank-k SVD see Fig. 1, the grey areas determine first k coordinates from singular vectors, which are being used.

$$\begin{bmatrix} A_k \\ n \times m \end{bmatrix} = \begin{bmatrix} U_k \\ n \times k \end{bmatrix} \begin{bmatrix} \Sigma_k \\ k \times k \end{bmatrix} \begin{bmatrix} V_k^T \\ k \times m \end{bmatrix}$$

Fig. 1 k-reduced singular value decomposition

**Theorem:** Among all  $m \times n$  matrices C of rank at most k,  $A_k$  is the one that minimizes

$$\|A_k - A\|_F^2 = \sum_{i,j} (A_{i,j} - C_{w,i,j})^2$$

Because rank-k SVD is the best rank-k approximation of original matrix A, any other decomposition will increase the sum of squares of matrix  $A - A_k$ .

The SVD is hard to compute and once computed, it reflects only the decomposition of the original matrix. The recalculation of SVD is expensive, so it

is impossible to recalculate SVD every time new rows or columns are inserted. The SVD-Updating is a partial solution, but since the error increases slightly with inserted rows and columns when updates occur frequently, the recalculation of SVD may be needed.[3][4][5][6].

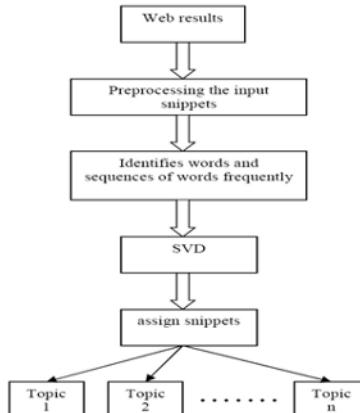
## 4 Problem Formalization and Algorithm

The distinctive characteristic of the algorithm is that it identifies meaningful cluster labels and only then assigns search results to these labels to build proper clusters. The algorithm consists of five steps:

1. Pre-processing the input snippets, which includes tokenization, stemming and stop-word marking.
2. Identifies words and sequences of words frequently appearing in the input snippets.
3. Matrix decomposition is used to induce cluster labels.
4. Snippets are assigned to each of these labels to form proper clusters.
5. Post-processing, which includes cluster merging and pruning.

The step 3 is the core of the algorithm, because this step relies on the Vector Space Model and a term document matrix A having n rows, where n is the number of input snippets, and m columns, where m is the distinct words found in the input snippets. Each element  $A_{nm}$  of A numerically represents the relationship between word m and snippet n.

The singular value decomposition may be applied on the binary matrix A which created from the step 2, where the rows of the matrix are the input snippets (objects), and the columns are the distinct words found in the input snippets (attributes), presented as 0 and 1.(0 – the distinct word not found in the input snippet, 1 – the distinct word found in the input snippet).



**Fig.2** Steps of the algorithm process

Since the rank-k SVD is known to remove noise by ignoring small differences between row and column vectors of A (they will correspond to small singular values, which we drop by the choice of k), it can be used in our algorithm. Because SVD creates equivalence classes of data from the original data through deleting and adding some no primary attributes in the objects, this process leads the objects to have similarity in their attributes. From this similarity the algorithm combines these objects that have the same attributes and presents them like one object.

The new object created from this process represent a group of snippets has a similar distinct word. That means that the algorithm can minimize the huge volume of snippets received as a result from the searching in the web.

## 5 Experiment

We applied our experiment on data created from Google search engine contained from 100 snippets with 35 distinct different words, but for easy explanation how the algorithm works, we deduct a small part from these data.

The data which we deducted from the result of Google search engine (table 1) are 20 objects (snippets) with 6 attributes (distinct words). These data have peculiar combination in every object.

**Table 1.** Data deducted from the result of Google search engine

	A1	A2	A3	A4	A5	A6
O1	0	0	0	0	0	1
O2	1	0	0	1	0	1
O3	0	0	0	1	0	1
O4	0	0	0	1	0	0
O5	1	0	0	1	0	0
O6	1	0	0	0	0	0
O7	0	1	1	1	1	0
O8	0	0	0	1	1	0
O9	0	1	1	0	0	0
O10	0	1	0	1	1	0
O11	0	0	1	1	1	0
O12	0	1	1	0	1	0
O13	0	1	1	1	0	0
O14	0	1	0	1	0	0
O15	0	1	0	0	1	0
O16	0	0	1	1	0	0
O17	0	0	1	0	1	0
O18	1	0	0	1	1	0
O19	0	1	1	0	0	1
O20	1	1	1	0	0	1

After using SVD with good choice k-rank, we can see that the output data from the SVD method are changed and brewed new attributes combinations in the objects, these combinations repeated in more than one object (table 2).

**Table 2.** Data after using SVD with Kruskal

	A1	A2	A3	A4	A5	A6
O1	0	0	0	0	0	0
O2	1	0	0	1	0	0
O3	1	0	0	0	0	1
O4	1	0	0	0	0	1
O5	1	0	0	0	1	0
O6	0	0	0	0	1	0
O7	0	1	1	1	1	0
O8	1	0	0	0	1	0
O9	0	1	1	1	0	0
O10	0	1	1	1	1	0
O11	0	1	1	1	1	0
O12	0	1	1	1	0	0
O13	0	1	1	1	0	0
O14	0	1	1	1	0	0
O15	0	0	0	1	0	0
O16	1	1	1	1	0	1
O17	0	0	0	1	0	0
O18	0	0	0	0	1	0
O19	0	1	1	1	0	0
O20	0	1	1	1	1	0

From this repeating of attributes combination in the objects, not imperative to represent all the objects with the same attributes, we can only represent the objects that have peculiar combinations of attributes (table 3).

From the above table, we can see how these attributes changed to have similar attributes combination. And we can also see that we have a more active change when the object has more attributes.

This process minimizes the representing data to 40% from the original data with saving the primary attributes in the objects.

**Table 3.** Representing data after combining the similar objects

	A1	A2	A3	A4	A5	A6
O1	0	0	0	0	0	0
O2	1	0	0	1	0	0
O3	1	0	0	0	0	1
O4	1	0	0	0	1	0
O5	0	0	0	0	1	0
O6	0	1	1	1	1	0
O7	0	1	1	1	0	0
O8	0	0	0	1	0	0

The activity of this method is shown clearly on the huge data with big quantity of attributes, and for a good choice k-rank from the diagonal matrix when applying SVD. Figure 3 showed how the k-rank value plays role in the number of attributes combination.

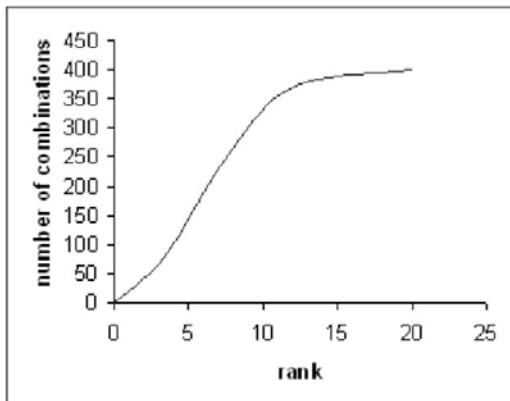


Fig. 3. Relationship between value of k-rank and number of combination

## 6 Conclusion and Future Work

Singular Value Decomposition (SVD) with k-rank depends on deleting and adding some attributes from the objects in the original data. This way can give us minimum primary attributes to collect more objects that have the same combination of attributes. Applying this method of decomposition on web searching problems gives us a good solution for search results clustering.

## Reference

1. Osinski, S. :Improving Quality of Search Results Clustering with Approximate Matrix Factorisations. ECIR (2006)
2. Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, Jinwen Ma : Learning to cluster web search results. SIGIR (2004)
3. Snasel V., Gajdos P., Abdulla H.D., Polovincak M. :Concept Lattice Reduction b y Matrix Decompositions. DCCA (2007).
4. Snasel V., Gajdos P., Abdulla H.D., Polovincak M. : Behavior of the Concept Lattice Reduction to visualizing data after Using Matrix Decompositions. IEEE Innovations'07, (2007)
5. Snasel V., Polovincak M., Abdulla H.D., Horak Z. : On Knowledge Structures Reduction. IEEE CISIM (2008)

6. Snasel V., Polovincak M., Abdulla H.D., Horak Z. : On Concept Lattices and Implication Bases from Reduced Contexts. ICCS Supplement (2008)
7. Berry M. , Browne M., : Understanding Search Engines, Mathematical Modelling and Text Retrieval. Siam, (1999)
8. Berry M., Dumais S., Letsche T. : Computation Methods for Intelligent Information Access. In Proceedings of the 1995 ACM/IEEE Supercomputing Conference (1995)
9. Larsen R.M. : Lanczos bidiagonalization with partial reorthogonalization. Technical report, University of Aarhus (1998)

# **Automatic Performance Evaluation of Web Search Systems using Rough Set based Rank Aggregation**

Rashid Ali<sup>1</sup> and M. M. Sufyan Beg <sup>2</sup>

<sup>1</sup> Department of Computer Engineering, A.M. U., Aligarh-202002, India  
rashidalamu@rediffmail.com

<sup>2</sup> Department of Computer Engineering, J.M.I., New Delhi- 25, India  
mmsbeg@hotmail.com

**Abstract.** Web searching is such an activity that its importance can just not be ignored in the current scenario. Since there are a large number of publicly accessible search engines, which differ in their indexing algorithms and hence the search results, the evaluation of search engines performance is needed to determine which one is the best. The human intelligence may be used to measure the search engine effectiveness. But, a subjective evaluation done on the basis of user-feedback is costly in terms of the time required. Therefore, it is also not scalable. So, there is a need of an automatic evaluation method. In this paper, we present the architecture of an automatic Web search evaluation system that combines the different evaluation techniques using a Rough Set based Rank aggregation technique. The rough set based rank aggregation models the user's feedback based rank aggregation. In the rough set based aggregation technique, the ranking rules are learnt on the basis of the user feedback in the training data sets. The learned rules are then used to estimate the overall ranking for the other data sets, for which user feedback is not available. We show our experimental results pertaining to seven public search engines.

## **1 Introduction**

Web Searching is arguably the second most popular activity on Internet. A number of public search engines are available for this purpose. In the Web based searching, a user queries the search engine for some query and gets the results in some order. Since different search engines employ different search algorithms and indexing techniques, the user gets different results in different order in response to the same query. So naturally, the quality of search comes into picture. A general user is just interested in viewing the results on the first page and in that too, just the top few only. So, the search engine, which gives the results important for the user from amongst the top few ones, should be voted as a better one. For this, the search results need to be evaluated. Such an evaluation helps in identifying the

most effective one and helps the users to find the required information with less effort. Such a study is needed at the personal as well as the business level. Moreover, due to the changing needs of users or the dynamic nature of search engines (e.g. their changing web coverage and ranking technology), performance evaluation of Web search engines is needed to be done on a regular basis. Hence, the evaluation procedure should be an efficient one. Besides helping the user in satisfying their information need with least effort, the evaluation is also poised to motivate search engine providers for higher standards. Moreover, based on this evaluation study, some new and efficient meta-search engines can also be developed. But the problem is how to evaluate the search engines.

Search engines can be evaluated subjectively on the basis of user feedback. The user feedback may be explicit or implicit in nature. An explicit feedback is the one in which the user is asked to fill up a feedback form after he has finished searching. This form is easy to analyze as the user may be asked directly to rank the documents as per the relevance according to his evaluation. The problem with the form-based approach is that it is too demanding for the user. In this approach, there is a lot of work for a casual user who might either fill it carelessly or not fill it at all. Therefore, there is a need to obtain the implicit feedback from the users. The feedback is implicit if we infer the feedback from the user by watching the actions of the user on search results presented before him in response to his query.

The problem with the user feedback based method is that it is a time costly affair. Therefore, it is also not scalable, that means it can be performed with a small amount of data but not with a very large amount of data. For larger amount of data, we cannot bear the time required in obtaining the user feedback. For larger data sets, automatic evaluation is the only answer. Automatic evaluation of Web search systems is desirable to perform the evaluation with a large number of queries and also for the evaluation, which is needed to be done frequently. On the other hand, usefulness of real user's judgment is also well-known. Therefore, there is a need to devise an efficient method for the automatic evaluation that incorporates the real user's judgment in some way. Different objective techniques like Vector Space Model [1], Boolean Similarity Measures [2], etc. can be used to evaluate the search engines automatically. But, automatic evaluation with each of these techniques lacks user's intelligence in addition to the individual limitations of the particular technique. Therefore, in this work, we propose to aggregate these techniques using rough set based rank aggregation technique. In the rough set based rank aggregation technique [3], the different objective evaluation techniques are combined on the basis of the user's feedback. The user feedback is obtained implicitly by watching the actions of the user on the search results in response to the queries in the training set. In other words, the rank aggregation technique exploits human intelligence in learning the ranking rules using rough set theory. Since, user's preference is a costly affair; the rough set based approach is very much useful because it models the user feedback based aggregation without user's involvement, once the system is trained.

This paper is organized as follows. In section 2, we look at the background and the related work. We then discuss the basics of the rough set theory and rough set based rank aggregation and present the architecture of the automatic Web search

evaluation system in the section 3. We present our experimental results in section 4. Finally, we conclude in section 5.

## 2 Background and Related Work

Let us begin with discussion on the related work in the area.

### 2.1 Related Work

In the past, few attempts have been made for the automatic evaluation of the Web search systems [4–12]. Soboroff et al. made the first evaluation study, in which, relevance judgment was performed automatically [4]. Soboroff et al. proposed to replace human relevance judgments with a number of randomly selected “pseudo-relevant” documents from a pool generated in the TREC environment. In his work, the contents of the documents were not considered for relevance judgments. Chowdhury and Soboroff presented a method for comparing search engine performance automatically based on how they rank the known item search result [5]. In their study, they constructed a large number of query document pairs. Queries were mined from search service query logs and documents were mined from the Open Directory Project [13]. Shang and Li evaluated six popular search engines namely AltaVista, Fast, Google, Go, iWon, and Northern Light, with 3000 queries from two domains, using a largely automatic test design [6]. They computed relevance scores using three difference relevance algorithms and statistical comparisons of the ranking. Wu and Crestani used the reference count method for automatic ranking of retrieval systems [7]. In their method, for each query, they first considered the list of documents returned by a retrieval system, and then noted references for each document of the list. Specifically, they counted the occurrences of the document in the lists provided by other systems. They considered each list one by one, performed the summation of these reference counts, and rank documents using the total reference count sum for each document. Can et al. proposed an automatic performance evaluation method for Web search engine evaluation [8]. They presented Automatic Web Search Engine Evaluation method (AWSEEM) as an efficient and effective evaluation tool for Web search systems. In their work, they proposed to replace human-based relevance judgments with a set of automatically generated relevance judgments. In their experiments, for each query, the top 200 documents from each search engine were collected to form a pool of documents. Then these documents were indexed and ranked using the Vector Space Model and were sorted in descending order according to their similarity to the query. Top 20 documents in this ranking were treated as Pseudorels. Can et al. evaluated the performance of search engines by considering these pseudo-(or automatic) relevance judgments as different human relevance judgments. Using Pearson’s r correlation, the researchers showed that their method provided results consistent with human-based evaluations. Beitzel et

al. also adopted a method similar to the AWSEEM method and used the Open Directory Project categories to determine (pseudo) relevant documents for the evaluation of Web search engines [9]. Sharma and Jansen discussed the development of a system that evaluated a search engine's performance using implicit user feedback in the real-time [10]. Relevant judgment data for performance evaluation was collected using implicit feedback by user in terms of his actions on search results. Aslam et al. developed a sampling approach in order to replace human relevance judgment [11]. Nuray and Can discussed new methods for automatic ranking of retrieval systems [12]. In their approach, they merged the retrieval results of multiple systems using various data fusion algorithms, use the top-ranked documents in the merged result as the “pseudo-relevant documents”, and used these documents to evaluate and rank the systems.

In this paper, we discuss the automatic evaluation of Web search systems using the rough set based rank aggregation. But, firstly, we list a few important definitions.

## 2.2 Important Definitions

**Definition 1.** Given a universe  $U$  and  $S \subseteq U$ , an ordered list (or simply, a list)  $l$  with respect to  $U$  is given as  $l = [e_1, e_2, \dots, e_{|S|}]$ , with each  $e_i \in S$ , and  $e_1 > e_2 > \dots > e_{|S|}$ , where “ $>$ ” is some ordering relation on  $S$ . Also, for  $j \in U \wedge j \in l$ , let  $l(j)$  denote the position or rank of  $j$ , with a higher rank having a lower numbered position in the list. We may assign a unique identifier to each element in  $U$  and thus, without loss of generality we may get  $U = \{1, 2, \dots, |U|\}$ .

**Definition 2. Full List:** If a list contains all the elements in  $U$ , then it is said to be a full list.

**Example 1.** A full list  $l_f$  given as  $[e, a, d, c, b]$  has the ordering relation  $e > a > d > c > b$ . The Universe  $U$  may be taken as  $\{1, 2, 3, 4, 5\}$  with say  $a \equiv 1, b \equiv 2, c \equiv 3, d \equiv 4, e \equiv 5$ . With such an assumption, we have  $l_f = [5, 1, 4, 3, 2]$ . Here  $l_f(5) \equiv l(e) = 1, l_f(1) \equiv l(a) = 2, l_f(4) \equiv l(d) = 3, l_f(3) \equiv l_f(c) = 4, l_f(2) \equiv l_f(b) = 5$ .

**Definition 3. Partial List:** A list  $l_p$  containing elements, which are a strict subset of universe  $U$ , is called a partial list. We have a strict inequality  $|l_p| < |U|$ .

**Definition 4. Spearman Rank Order Correlation coefficient** [14]: Let the full lists  $[u_1, u_2, \dots, u_n]$  and  $[v_1, v_2, \dots, v_n]$  be the two rankings for some query  $Q$ . Spearman rank-order correlation coefficient ( $r_s$ ) between these two rankings is defined as follows-

$$r_s = 1 - \frac{6 \sum_{i=1}^n [l_f(u_i) - l_f(v_i)]^2}{n(n^2 - 1)} \quad (1)$$

The Spearman rank-order correlation coefficient ( $r_s$ ) is a measure of closeness of two rankings. The coefficient  $r_s$  ranges between  $-1$  and  $1$ . When the two rankings are identical  $r_s = 1$ , and when one of the rankings is the inverse of the other then the  $r_s = -1$ .

**Definition 5. Modified Spearman Rank Order Correlation coefficient:** Without loss of generality, assume that the full list is given as  $[1, 2, \dots, n]$ . Let the partial list be given as  $[v_1, v_2, \dots, v_m]$ . The Modified Spearman rank-order correlation coefficient ( $r_s'$ ) between these two rankings is defined as follows-

$$r_s' = 1 - \frac{\sum_{i=1}^m (i - v_i)^2}{m \left( \left[ \max_{j=1}^m \{v_j\} \right]^2 - 1 \right)} \quad (2)$$

**Example 2.** For  $|U|=5$ , let the full list be  $l_f = \{1, 2, 3, 4, 5\}$  and the partial list  $l_p$  with  $|l_p| = m = 3$  be  $l_p = \{40, 35, 100\}$ .

$$r_s' = 1 - \frac{(1 - 40)^2 + (2 - 35)^2 + (3 - 100)^2}{3 \times \left( \left[ \max \{40, 35, 100\} \right]^2 - 1 \right)} = 0.401$$

**Definition 6. Rank Aggregation Problem:** Given a set of  $n$  candidates say,  $C = (C_1, C_2, \dots, C_n)$ , a set of  $m$  voters, say  $V = (V_1, V_2, \dots, V_m)$ , and a ranked list  $l_i$  on  $C$  for each voter  $i$ . Then,  $l_i(j) < l_i(k)$  indicates that the voter  $i$  prefers the candidate  $j$  to  $k$ . The rank aggregation problem is to combine the  $m$  ranked lists  $l_1, l_2, \dots, l_m$  into a single list of candidates say  $l$  that represents the “collective choice” of the voters. The function used to get  $l$  from  $l_1, l_2, \dots, l_m$  (i.e.  $f(l_1, l_2, \dots, l_m) = l$ ) is known as rank aggregation function.

### 3 An Automatic Web Search Evaluation System

In this section, we first discuss briefly about rough set theory and then present the details of the rough set based rank aggregation technique and then present the architecture of an automatic Web search evaluation system that uses rough set based rank aggregation to combine different evaluation techniques.

### 3.1 Rough Set Theory

Rough set theory, proposed by Pawlak in 1982[15], is a novel mathematical approach to vagueness. Rough set philosophy is based on the assumption that, in contrast to the classical set theory, we have some additional information (knowledge, data) about elements of a set. In rough set approach, it is assumed that a pair of precise concepts – called the lower and the upper approximation of the vague concept, replaces any vague concept. Therefore, for each rough set, two crisp sets, called the lower and the upper approximation of the rough set, are associated. The lower approximation consists of all objects, which surely belong to the set, and the upper approximation contains all objects, which possibly belong to the set. The difference between the upper and the lower approximation constitute the boundary region of the rough set. Hence, rough set theory expresses vagueness by employing a boundary region of a set. The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information about data like a grade of membership or the value of possibility in fuzzy set theory, probability distributions in statistics etc.

For data analysis, objects are generally represented in terms of their values on a set of attributes. Specifically, to present specific information about objects notion of an information table is used. An information table is a two-dimensional structure where rows correspond to objects of the universe, the columns correspond to a set of attributes, and each cell is the value of an object with respect to an attribute. Formally, an information table is a quadruple  $T = (U, A_t, \{V_a | a \in A_t\}, \{I_a | a \in A_t\})$

Where,

$U$  is a finite nonempty set of objects.

$A_t$  is a finite nonempty set of attributes.

$V_a$  is a nonempty set of values for  $a \in A_t$ .

$I_a : U \rightarrow V_a$  is an information function.

When, we are dealing with a special attribute  $d$  called decision attribute, then the information table is specifically called decision table and all attributes other than  $d$  are called conditional attributes. To reduce the size of decision table, two tasks are performed on the table. The first is to identify equivalence classes i.e. the objects that are indiscernible using the available attributes. So that, only one element is needed to represent entire class. The second task is to search for attributes that preserve indiscernibility relation and consequently set approximation. The rest of the attributes are redundant. There are too many such subsets of attributes and those that are minimal are called *reducts*. The union of all *reducts* is called *core*. A good description of rough set theory and its application can be found in [16].

### 3.2 Rough Set-Based Rank Aggregation

In the rough set based technique, rank aggregation is performed in two phases. In the first phase, the system learns ranking rules using rough set theory from the

user feedback based ranking available in the training data. The user feedback based ranking is considered as the overall ranking for the purpose. The best set of learned ranking rules is then selected by performing cross validation. In the second phase, the best set of learned ranking rules is used to estimate the overall ranking for the set of rankings for which user feedback is not available. The estimated overall ranking is thus a model of user feedback based ranking and is presumed to be the best overall ranking.

Our method for learning ranking rules is similar to the method discussed in [17] for mining ordering rules. As in [17], we also convert our information table called ranked information table to binary information table. In the binary information table, an equivalence relation  $E_A$  for a subset of attributes  $A \subseteq A_t$  can be defined. The attribute corresponding to the overall ranking partitions all pairs of objects into two disjoint classes. The lower and upper approximation of each class can simultaneously be obtained based on attributes in  $A$ . *Reducts* and *Core* of attributes in  $A$  can also be found to eliminate the redundant ones. Then, for each equivalence class present in lower approximation class, a certain rule can be drawn. A possible rule can also be drawn from equivalence class present in upper approximation. These possible rules are useful in case of larger data sets where inconsistencies may reduce lower approximation and hence the finding of strong rules. Rosetta, a rough set toolkit for analyzing data [18], may be used to get a minimal set of ranking rules from the binary information table. The minimal set of the ranking rules thus obtained is validated using cross validation technique to select the best set of ranking rules.

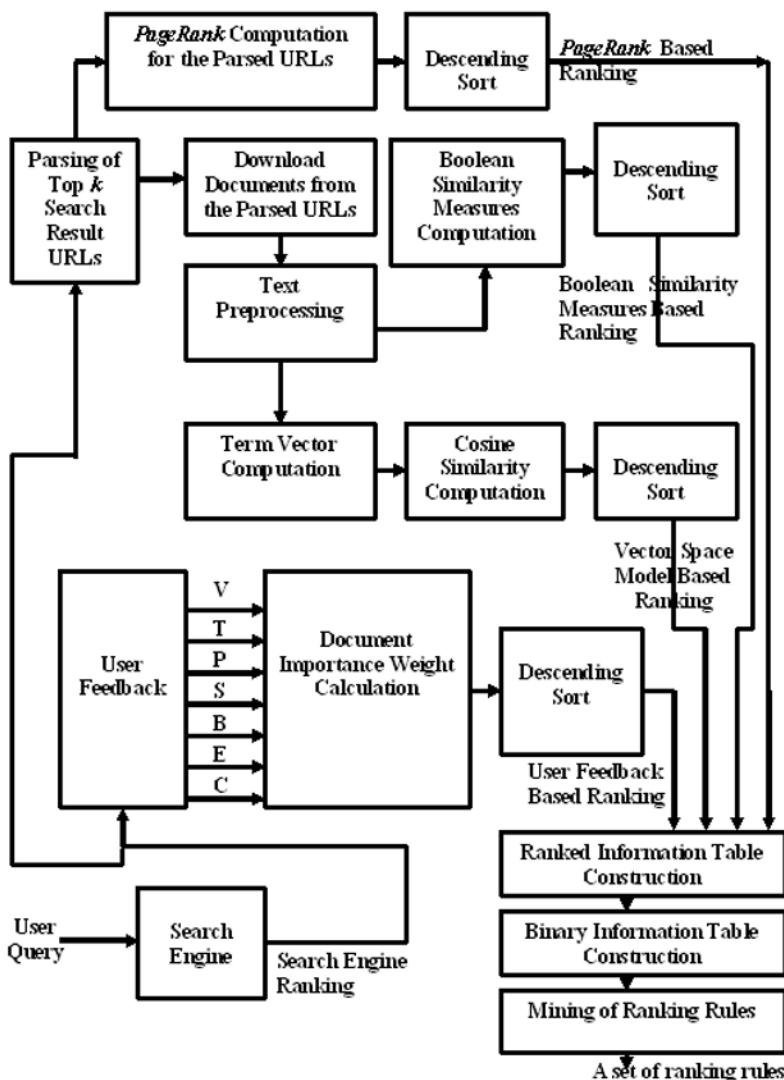
The selected ranking rules are used to predict user feedback based ranking for the future data. For this, again we follow the similar procedure as in the learning ranking rules. In this case, since we do not have user feedback based ranking, the values of the attribute corresponding to the user feedback based ranking in the ranked information table may be initialized as all zeroes. Then, we can estimate the values of the attribute corresponding to the user feedback in the binary information table by using the selected ranking rules. After that, we can convert the binary information table back into the ranked information table. The attribute corresponding to the user feedback based ranking in the ranked information table gives the predicted overall ranking.

### 3.3 Architecture of an Automatic Web Search Evaluation System

Now, we propose the architecture of an automatic Web search evaluation system, which evaluates the search systems automatically on the basis of rough set based aggregation. The evaluation system works in two phases namely ranking rules learning phase and rank aggregation phase. In the ranking rules learning phase, the user gives his query to the search engine and obtains the search results ordered by search engine ranking say  $R_{SE}$ . Then, the user feedback is taken implicitly by watching the actions of the user on the search results and is characterized by the

vector ( $V$ ,  $T$ ,  $P$ ,  $S$ ,  $B$ ,  $E$ ,  $C$ ) [19]. The three evaluation techniques based on Vector Space Model, Boolean Similarity measure and *PageRank* [20] are applied on the top  $k$  documents returned by the search engine. In the process, the four different rankings of the documents namely  $R_{VS}$ ,  $R_{BS}$ ,  $R_{PR}$  and  $R_{JUF}$  are obtained from the evaluations based on Vector Space Model, Boolean Similarity Measures, *PageRank* and implicit user feedback respectively. The ranked information table is constructed from these four rankings, which is then converted into the binary information table. Then, the ranking rules are mined using Rosetta. The process is repeated for a number of queries and the ranking rules obtained in the process are validated to select the set of ranking rules that will be used for the aggregation in the rank aggregation phase. The overall procedure for the learning phase is shown in Fig. 1.

In the rank aggregation phase, a query is issued to the search system and the top  $k$  search results returned in response to the query are evaluated using the three evaluation techniques based on Vector Space Model, Boolean Similarity Measures and *PageRank*. The three different rankings of the documents namely  $R_{VS}$ ,  $R_{BS}$ ,  $R_{PR}$  obtained in the process are used to construct the ranked information table, which is then converted into the binary information table. The ranking rules obtained from the ranking rules learning phase with their corresponding confidence values are then used to predict the column corresponding to the user feedback. On the basis of which a score is computed for each document as in the rank aggregation algorithm. Sorting the document on the descending order of their respective scores gives an overall ranking of the documents; say  $R_{all}$ , which is then compared with the search engine ranking, computed. The overall process is repeated for a fairly large number of queries and the  $r_s'$  obtained for each query is averaged. The averaged value of  $r_s'$  is then the quantitative measure, say RBSQM, of the search quality of the search engine. The process may be repeated for different search engines and the search engines can be graded on the basis of this quantitative measure. The overall procedure for the rank aggregation phase is shown in Fig. 2.



**Fig. 1.** Ranking Rules learning phase of the System

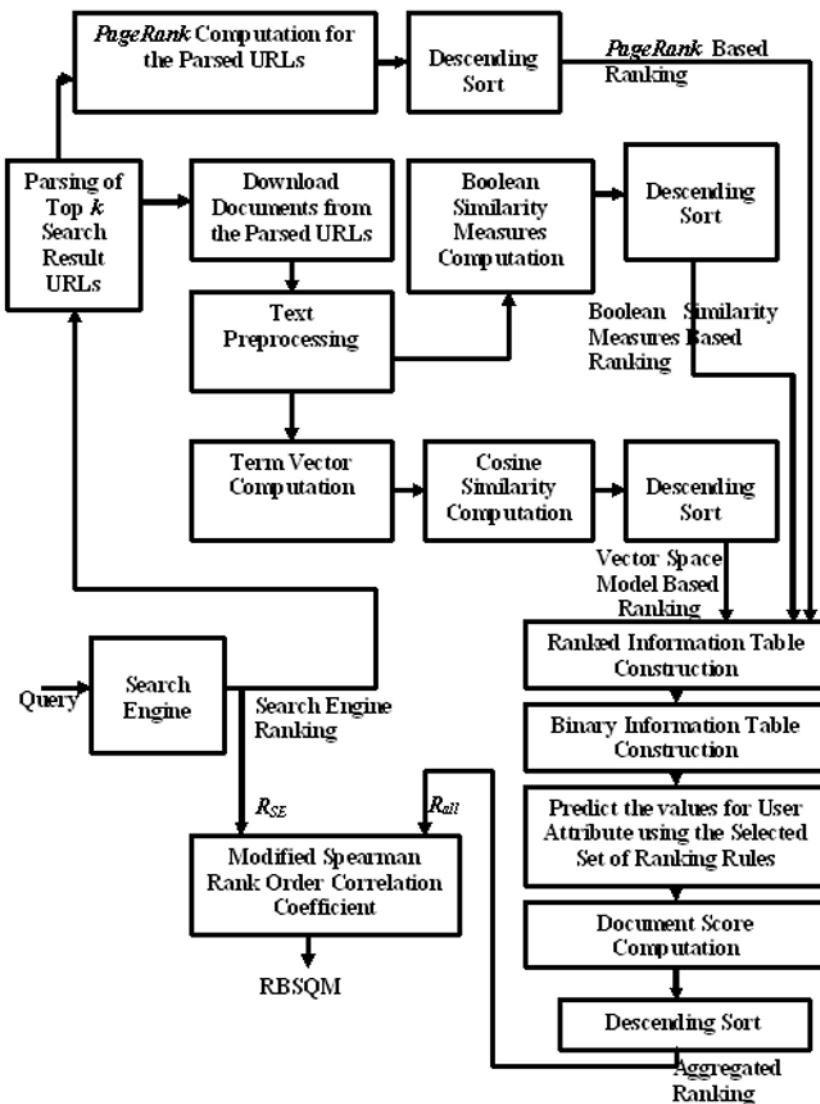


Fig. 2. Rank Aggregation phase of the System

## 4 Experiments and Results

We evaluated the search results of seven popular search engines, namely, *AltaVista*, *Ask*, *Excite*, *Google*, *HotBot*, *Lycos* and *Yahoo*. For the automatic performance evaluation of these Web search systems, we experimented with more than 500 queries. We constructed these queries from the keywords of research papers published in the twelve issues of IEEE Transactions on Knowledge and Data Engineering (from January 2006 to December 2006).

For learning the ranking rules, we used 15 queries. We used the same set of 15 queries, which was used in [19, 21], and is listed in Table 1. For each query in the learning phase, we presented the top 10 search results obtained in response to the search query before the user and obtained user feedback implicitly by watching actions of user on the search results. On the basis of the user feedback, the document importance weights were calculated and the documents were sorted in the descending order of their document importance weights to obtain user feedback based ranking say  $R_{IUF}$ . Similarly, we evaluated the top 10 search results using Vector Space Model based evaluation, Boolean Similarity Measures based evaluation and *PageRank* based evaluation and obtained the three different rankings of the documents namely  $R_{VS}$ ,  $R_{BS}$  and  $R_{PR}$  based on the Vector Space Model, Boolean Similarity Measures and *PageRank* respectively using the process shown in Fig. 1. With these four rankings  $R_{IUF}$ ,  $R_{VS}$ ,  $R_{BS}$  and  $R_{PR}$ , we constructed ranked information table. Since, user feedback ranking may be a partial list as user may select only a few of the top 10 documents, we place a value -11 for the documents not selected by the user in the ranked information table. We converted the ranked information table into binary information table. Then, using Rosetta, we mined the ranking rules. We repeated the process for each of the seven search engines and in the process obtained seven different set of ranking rules.

Since, the number of queries used for learning ranking rules is very less in comparison of total number of queries used in the aggregation phase, we considered each mined rules significant enough to be used as ranking rules in the aggregation phase. In other words, since training data is less, we use each and every training pattern for the classification purpose. Moreover, as the number of documents selected by user might be less, the class 0 may be dominant in the binary information table. Therefore, for prediction of the binary class, we take into account the confidence of prediction of 1 for the ranking rules that are not certainly predicting a single class.

**Table 1.** List of Queries used in Learning Phase

1	measuring search quality
2	mining access patterns from web logs
3	pattern discovery from web transactions
4	Distributed associations rule mining
5	document categorization query generation
6	term vector database
7	client -directory-server-model
8	similarity measure for resource discovery
9	hypertextual web search
10	IP routing in satellite networks
11	focussed web crawling
12	concept based relevance feedback for information retrieval
13	parallel sorting neural network
14	spearman rank order correlation coefficient
15	web search query benchmark

We validated the ranking rules by comparing the predicted user feedback based ranking with the actual user feedback based ranking. The results of the validation are presented in Table 2. From the Table 2, it is clear that there is a high correlation between the user feedback based ranking and the predicted user feedback based ranking for all the seven search engines. The highest average correlation is 0.86138 for *Altavista* and the lowest is 0.77628 for *Excite*. Then, using the selected set of ranking rules, we evaluated automatically the seven search engines in the rank aggregation phase. In the rank aggregation phase, for each query, we downloaded the top 10 documents from the search results and obtained the three different rankings  $R_{VS}$ ,  $R_{BS}$  and  $R_{PR}$  based on the Vector Space Model, Boolean similarity measures and *PageRank* respectively using the process shown in Fig. 2. Then we constructed the ranked information table using these three rankings and converted that into binary information table. Then, using the selected set of ranking rules from ranking rules learning phase, we predicted the column corresponding to the user feedback. Then, we computed the score for each of the 10 documents and sorted the documents in the decreasing order of their respective scores to obtain the aggregated ranking. The aggregated ranking was compared with the search engine ranking  $R_{SE}$  and the Modified Spearman Rank Order Correlation Coefficient ( $r_s'$ ) was computed.

We experimented with 543 queries in all. The correlation coefficients are averaged for all the 543 queries and the search engines are graded on the basis of the averaged value. The averaged values of correlation coefficients for all the 543

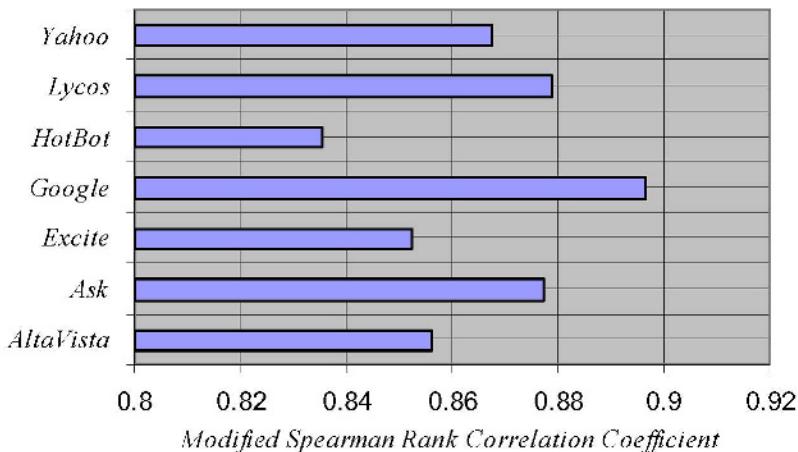
queries are listed in Table 3. The result of Table 3 is pictorially represented in Figure 3.

**Table 2.** Results of validation for the 15 queries

<i>Query</i>	<i>AltaVista</i>	<i>Ask</i>	<i>Excite</i>	<i>Google</i>	<i>HotBot</i>	<i>Lycos</i>	<i>Yahoo</i>
1	.73809	.95000	.88750	.73958	1	.79861	.81250
2	.85253	.75589	.81250	.79394	.82155	.97778	.90476
3	.75000	.93308	.78125	.93333	.85714	.87475	.96668
4	.85522	.80135	.89286	.90278	.72917	.88194	.83232
5	.85354	.92250	.81875	.78182	.91250	.88552	.88889
6	.90000	.82576	.93571	.89338	.72188	.63333	.89683
7	.90476	.93333	.82155	.97143	.81818	.89583	.28571
8	.85714	.78333	.72222	.74464	1	.76250	.91500
9	.98594	.60833	.68254	.98810	.72980	.91631	.71875
10	.82716	.86667	.85606	.77980	.90476	.89899	.77083
11	.92929	.86875	.59596	.86111	.64286	.76768	.77083
12	.89647	.82857	.63750	.88571	.61616	.71667	.77083
13	.88889	.90104	.62963	.51515	.28571	.77778	.77083
14	.78788	.77778	.92727	.79394	.65278	.78030	.77083
15	.89375	.73958	.64286	.62857	1	.93750	.77083
Average	<b>.86138</b>	<b>.83306</b>	<b>.77628</b>	<b>.81422</b>	<b>.77950</b>	<b>.83370</b>	<b>.78976</b>

**Table 3.** Average of Correlation Coefficient ( $r_s'$ ) for the 543 queries

	<i>AltaVista</i>	<i>Ask</i>	<i>Excite</i>	<i>Google</i>	<i>HotBot</i>	<i>Lycos</i>	<i>Yahoo</i>
Average	<b>.85633</b>	<b>.87739</b>	<b>.85240</b>	<b>.89661</b>	<b>.83548</b>	<b>.87878</b>	<b>.86748</b>



**Fig. 3.** Performance of Different Search Engines based on the Automatic Web Search Evaluation

From Table 3 and Figure 3, we observe that *Google* gives the best performance, followed by, *Lycos*, *Ask*, *Yahoo*, *AltaVista*, *Excite* and *HotBot*, in that order.

## 5 Conclusion

In this paper, we proposed the architecture of an automatic Web search evaluation system, which employs the different objective evaluation techniques and combines the different rankings of search results obtained from these techniques by using the rough set based rank aggregation. In our experiments, we used a set of 15 queries in the ranking rules learning phase. The selected sets of ranking rules were then used in the rank aggregation phase. Our results for a set of 543 queries and 7 public web search engines show that *Google* gives the best performance, followed by *Lycos*, *Ask*, *Yahoo*, *AltaVista*, *Excite* and *HotBot*, in that order.

## References

1. Salton, G., Wong, A., and Yang, C. S.: A vector space model for automatic indexing. Communications of the ACM 18, 613–620 (1975)
2. Li, S. H. and Danzig, P. B. Boolean similarity measures for resource discovery. IEEE Transactions on Knowledge and Data Engineering 9, 863–876 (1997)
3. Ali, R. and Beg, M. M. S.: Rough set based rank aggregation for the Web. In Proceedings of 3rd Indian International Conference on Artificial Intelligence (IICAI-07), Pune, India pp. 683–698 (2007)

4. Soboroff, I., Nicholas, C., and Cahan, P.: Ranking retrieval systems without relevance judgments. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New Orleans, LA, U.S.A. pp. 66–73 (2001)
5. Chowdhury, A. and Soboroff, I.: Automatic evaluation of World Wide Web search services. In Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, Tampere, Finland, ACM Press pp. 421–422 (2002)
6. Shang, Y. and Li, L.: Precision evaluation of search engines. *World Wide Web* 5, 159–173 (2002)
7. Wu, S. and Crestani, F.: Methods for ranking information retrieval systems without relevance judgments. In Proceedings of the ACM Symposium on Applied Computing, Melbourne, Florida, U.S.A. ) pp. 811–816 (2003)
8. Can, F., Nuray, R., and Sevdik, A. B.: Automatic performance evaluation of Web search engines. *Information Processing and Management* 40, 495–514 (2004)
9. Beitzel, S. M., Jensen, E. C., Chowdhury, A., and Grossman, D.: Using titles and category names from editor-driven taxonomies for automatic evaluation. In Proceedings of the 12th International Conference on Information and Knowledge Management, New Orleans, LA, U.S.A. pp. 17–23 (2003)
10. Sharma, H. and Jansen, B. J.: Automated evaluation of search engine performance via implicit user feedback. In Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval, Salvador, Brazil pp. 649–650 (2005)
11. Aslam, J. A., Pavlu, V., and Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, WA, U.S.A pp. 541–548 (2006)
12. Nuray, R. and Can, F.: Automatic ranking of information retrieval systems using data fusion. *Information Processing and Management* 42, 595–614 (2006)
13. Open Directory Project. <http://dmoz.org/>
14. Weisstein, E. W.: Spearman Rank Correlation Coefficient. From MathWorld – A Wolfram Web Resource, ©1999–2004 Wolfram Research, Inc. <http://mathworld.wolfram.com/SpearmanRankCorrelationCoefficient.html>
15. Pawlak, Z.: Rough sets. *International Journal of Computer and Information Sciences* 11, 341–356 (1982)
16. Komorowski, J., Pawlak, Z., Polkowski, L, Skowron, A.: Rough sets: A tutorial In Rough Fuzzy Hybridization: A New Trend in Decision-Making, S.K. Pal, A. Skowron (Eds.), Springer-Verlag, Singapore pp. 3–98 (1999)
17. Yao, Y.Y., Sai, Y.: Mining ordering rules using rough set theory. [J] *Bulletin of International Rough Set Society* pp. 599–106 (2001)
18. Rosetta, a rough set toolkit for analyzing data.  
<http://www.idi.ntnu.no/aleks/rosetta/>
19. Beg, M. M. S.: A subjective measure of Web search quality. *International Journal of Information Sciences*. 169, 365–381. (2005)
20. Page, L., Brin, S., Motwani, R., and Winograd, T.: The *PageRank* citation ranking: Bringing order to the Web. Technical report, Computer Science Department, Stanford University, U.S.A. (1999)
21. Beg, M. M. S. and Ahmad, N.: Web search enhancement by mining user actions. *International Journal of Information Sciences*. 177, 5203–5218 (2007)

# Knowledge Discovery and Data Mining

# **Minimizing Space Time Complexity by RSTDB a New Method for Frequent Pattern Mining**

Vaibhav Kant Singh<sup>1</sup> and Vinay Kumar Singh<sup>2</sup>

<sup>1</sup>Department of Computer Science & Engineering SATI Vidisha vibhu200427@gmail.com,

<sup>2</sup>Department of MCA GGU Bilaspur vks\_123123@rediffmail.com

**Abstract.** Data-mining is the extraction of meaningful patterns from the large source of data. Association Rule Mining (ARM) is an important data mining technique. Mining of frequent patterns is a very important association rule mining problem. The previous approach i.e. Apriori suffers from the candidate-generation and test mechanism. The Apriori approach becomes inefficient when either the length of the frequent set or length of the Transaction Database (TDB) increases. The algorithm adopts bottom up breadth first approach for the mining purpose. In this research work, we have proposed a Reduced Scanning Transaction Database (RSTDB) algorithm that uses certain heuristic function which reduces the number of Transaction Database passes required to generate the maximum frequent set required for Association Rule Mining (ARM). The approach is a hybrid of bottom up and top down approach. It uses both upward and downward closure properties for frequent item sets evaluation.

In this work we will compare the Apriori approach with the above proposed approach for frequent pattern mining. We will try to evaluate the shortcoming of the proposed approach and also look as to how much efficient it is and in which cases. The RSTDB algorithm not only reduces the database scans but also will help in reducing the number of candidate-generation for a phase that is having a value less than the minimum support threshold value.

## **1 Introduction**

The progress in digital data acquisition and storage has resulted in the growth of huge database. Data mining (or data discovery) is the process of autonomously extracting useful information or knowledge from large data stores or sets. Data mining typically deals with data that have already been collected for some purpose other than data mining analysis.

Association rule mining is a very popular data mining technique and it finds relationships among the different entities of records. Since the introduction of

frequent itemsets, it has received a great deal of attention in the field of knowledge discovery and data mining.

One of the first algorithms proposed for association rules mining was the AIS algorithm [1]. The problem of association rules mining was introduced in [1] as well. This algorithm was improved later to obtain the Apriori algorithm [2]. The Apriori algorithm employs the downward closure property if an item set is not frequent, any superset of it cannot be frequent either. The Apriori algorithm performs a breadth-first search in the search space by generating candidate  $k+1$ -itemsets from frequent  $k$  itemsets. The frequency of an item set is computed by counting its occurrence in each transaction.

FP-growth [3] is a well-known algorithm that uses the FP-tree data structure to achieve a condensed representation of the database transactions and employs a divide and-conquer approach to decompose the mining problem into a set of smaller problems. In essence, it mines all the frequent itemsets by recursively finding all frequent itemsets in the conditional pattern base that is efficiently constructed with the help of a node link structure. A variant of FP-growth is the H-mine algorithm [4]. It uses array-based and trie-based data structures to deal with sparse and dense datasets respectively. PatriciaMine [5] employs a compressed Patricia trie to store the datasets. FPgrowth\* [6] uses an array technique to reduce the FP-tree traversal time. In FP-growth based algorithms, recursive construction of the FP-tree affects the algorithm's performance.

Eclat [8] is the first algorithm to find frequent patterns by a depth-first search and it has been shown to perform well. It uses a vertical database representation and counts the item set supports using the intersection of tids. However, because of the depth-first search, pruning used in the Apriori algorithm is not applicable during the candidate itemsets generation. VIPER [9] and Mafia [10] also use the vertical database layout and the intersection to achieve a good performance. The only difference is that they use the compressed bitmaps to represent the transaction list of each item set. However, their compression scheme has limitations especially when tids are uniformly distributed. The dEclat [11] uses the vertical database representation. They store the difference of tids called diffset between a candidate  $k$  item set and its prefix  $k-1$  frequent itemsets, instead of the tids intersection set. They compute the support by subtracting the cardinality of diffset from the support of its prefix  $k-1$  frequent item set. This algorithm has been shown to gain significant performance improvements over Eclat[8]. However, when the database is sparse, diffset will lose its advantage over tidset.

## 1.1 Datamining

Data mining is the exploration and analysis of large datasets, in order to discover meaningful patterns and rules. Data mining is a component of a wider Process called (KDD) Knowledge Discovery from Database. Before a data set is mined, it first has to be cleaned. This removes, errors, ensures consistency and

takes missing values into account. Data mining may use quite simple or highly sophisticated data analysis. Data mining is a component of Data warehousing but it can be a stand alone process for data analysis, even in the absence of a Data warehouse.

Data mining is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining, the extraction of the hidden predictive information from large databases, is a powerful new technology with great potential to analyze important information in a Data ware-house.

A Data ware-house is a subject oriented, integrated, time-varying, non-volatile collection of data in support of the management's decision-making process.

Data warehousing is the Process of integrating enterprise-wide corporate data into a single repository, from which end-users can easily run queries, make reports and perform analysis. Data warehousing is a new approach to Enterprise-wide computing at the strategic or Architectural level.

## 1.2 Data Mining V/S Knowledge Discovery from Database

The Data mining process can be coupled with a DBMS in tightly coupled mode or loosely coupled mode. Data mining techniques support automatic exploration of data. Data mining attempts to source out trends or patterns in the data and infers rules from these patterns.

The term "Data Mining" refers to the finding of relevant and useful information from database. Data mining and knowledge discovery in the database is a new interdisciplinary field, merging ideas from statistics, machine learning and parallel computing.

Data Mining is only one of the many steps involved in the database. The various steps involved in KDD process include data selection, data cleaning and Preprocessing, data transformation and reduction, Data Mining Algorithm selection and finally the Post-Processing and the interpretation of the discovered knowledge. The KDD process tends to be highly iterative and interactive under computation:-

- KDD is the Process of identifying a valid, potentially useful and ultimately understandable structure in data.
- The structures that are outcomes of the data mining process must meet certain conditions to be considered as knowledge. These are validity, understandability, utility, novelty and interestingness.

### 1.3 Data Mining Techniques

The two fundamental goals of Data-Mining are:-

1. Prediction
2. Description

#### PREDICTION

Prediction makes use of the existing variables in the database in order to predict unknown or future values of interest.

#### DESCRIPTION

Description focuses on finding patterns describing the data and the subsequent presentation for user interpretation.

There are Several Data-Mining techniques fulfilling to the above goals:-

- **Association** :- The Presence of one set of items in a transaction implies other set of items
- **Classification:** - Develops profiles of different groups.
- **Sequential Patterns:** - Identifies sequential patterns subject to user constraints.
- **Clustering:** - Segments database into subsets or clusters.

## 2 The Apriori Algorithm

**Apriori** is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent item set properties.

Apriori employs an iterative approach known as a level wise search, where k-itemsets are used to explore (k+1) itemsets. First, the set of frequent 1 itemsets is found. This set is denoted  $L_1$ .  $L_1$  is used to find  $L_2$ , the set of frequent 2-itemsets , which is used to find  $L_3$ , and so on , until no more frequent k-itemsets can be found . The finding of each  $L_k$  requires one full scan of database. To improve the efficiency of the level – wise generation of frequent itemsets, an important property called the Apriori property, is used to reduce the search space. In order to use the Apriori property, all nonempty subsets of a frequent item set must also be frequent. This property is based on the following observation. By definition , if an item set  $I$  does not satisfy the minimum support threshold ,  $\text{min\_sup}$  , then  $I$  is not frequent , that is ,  $P(I) < \text{min\_sup}$ . If an item  $A$  is added to the item set  $I$ , then the resulting item set (i.e.  $I \cup A$ ) can not occur more frequently than  $I$ . Therefore,  $I \cup A$  is not frequent either, that is  $P(I \cup A) < \text{min\_sup}$ .

There are two steps for understanding that how  $L_{k-1}$  is used to find  $L_k$ . :-

1. The join step :-

To find  $L_k$ , a set of candidate k-itemsets is generated by joining  $L_{k-1}$  with itself. This set of candidates is denoted  $C_k$ .

2. The prune step:-

$C_k$  is a superset of  $L_k$ , that is , its members may or may not be frequent , but all of the frequent k-itemsets are included in  $C_k$  .

A scan of the database to determine the count of each candidate in  $C_k$  would result in the determination of  $L_k \cup C_k$ , however, can be huge , and so this could involve heavy computation . To reduce the size of  $C_k$  , the Apriori property is used as follows .

Any  $(k-1)$ -item set that is not frequent cannot be a subset of frequent k-item set.

Hence, if  $(k-1)$  -subset of a candidate k- item set is not in  $L_{k-1}$  , then the candidate cannot be frequent either and so can be removed from  $C_k$ .

## 2.1 PROBLEM DOMAIN

The Major Computational Challenges faced by Apriori are multiple scans of Transaction Database, Huge number of Candidates and Tedious workload of support counting for Candidates. For improving Apriori the general ideas that should be incorporated are to reduce passes of transaction database scans, shrinking the number of candidates and facilitating support counting of candidates.

## 3 Related Works

### 3.1 Previous Approach

The FP-growth algorithm [3] is one of the fastest approaches for frequent item set mining. The FP-growth algorithm [3] uses the FP-tree data structure to achieve a condensed representation of the database transaction and employs a divide-and-conquer approach to decompose the mining problem into a set of smaller problems. In essence, it mines all the frequent itemsets by recursively finding all frequent itemsets in the conditional pattern base that is efficiently constructed with the help of a node link structure.

The Mafia algorithm [10] uses vertical database layout and does intersection to achieve good performance. The search strategy of the algorithm integrates a depth-first traversal of the item set lattice with effective pruning mechanisms that significantly improve mining performance. The Mafia algorithm [10] uses vertical

database layout and intersection .It uses compressed bitmaps to represent the transaction list of each item set.

The dEclat algorithm [11] makes use of the vertical database representation where each item maintains a set of transaction ids where this item is contained. They store the difference of ids, called the diffset, between the candidate item set and its prefix frequent item sets, instead of the tids intersection set. They compute the support by subtracting the cardinality of diffset from the support of its prefix frequent item set.

### 3.2 RSTDB Algorithm

In this paper I am proposing an algorithm called RSTDB which reduces the number of scans involved in Apriori for this we will be using heuristic function which calculates the overall number of times the scanning is going to be done before actually iteration starts this reduces the number of passes required for frequent set estimation.

#### 3.2.1 RSTDB Algorithm:-

➤ **STEP 1:**

Calculate the size of each transaction in the Transaction Database.

➤ **STEP 2:**

Evaluate the transaction set having maximum size.

➤ **STEP 3:**

Check for the Transaction set size having frequency or support value more than the given threshold value. Set this Transaction size as the maximum value up to which scanning & candidate-generation step has to proceed. This will be the maximum value of k up to which iteration will be done

➤ **STEP 4:**

Candidate-Generation

gen\_cand\_itemsets with the given  $L_{k-1}$  as follows

$$C_k = \phi \dots \text{Equation(3.1)}$$

For all item set  $l_1 \in L_{k-1}$  do

For all item set  $l_2 \in L_{k-1}$  do

If  $l_1[1]=l_2[1] \wedge l_1[2]=l_2[2] \wedge \dots \wedge l_1[k-1] < l_2[k-1]$

Then  $c = l_1[1], l_1[2], \dots, l_1[k-1], l_2[k-1]$

$$C_k = C_k \cup \{c\} \dots \text{Equation(3.2)}$$

➤ **STEP 5:**

Candidate Set Pruning

Prune( $C_k$ )

For all  $c \in C_k$

For all ( $k-1$ ) subsets  $d$  of  $c$  do

If  $d \notin L_{k-1}$

Then

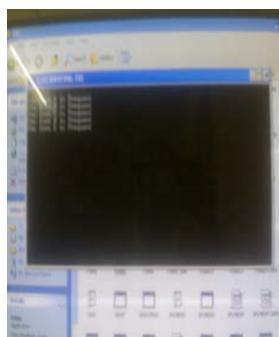
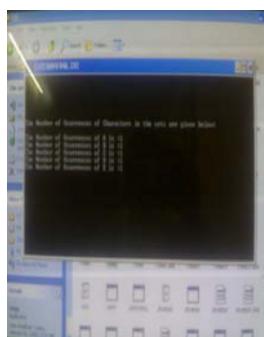
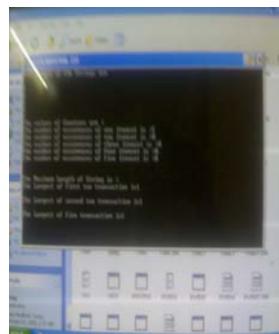
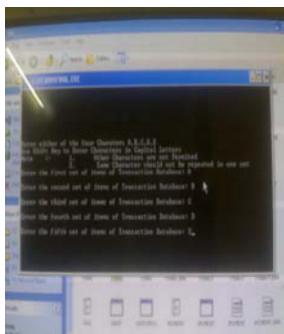
$$C_k = C_{k-1} \setminus \{c\} \dots \dots \text{Equation( 3.3 )}$$

Here,

$k$  is the number of passes required.

$L_{k-1}$  is the frequent item set.

$C_k$  is the candidate item set.



**Fig. 1.** The Output Windows in C++ of RSTDB Algorithm showing working of RSTDB.

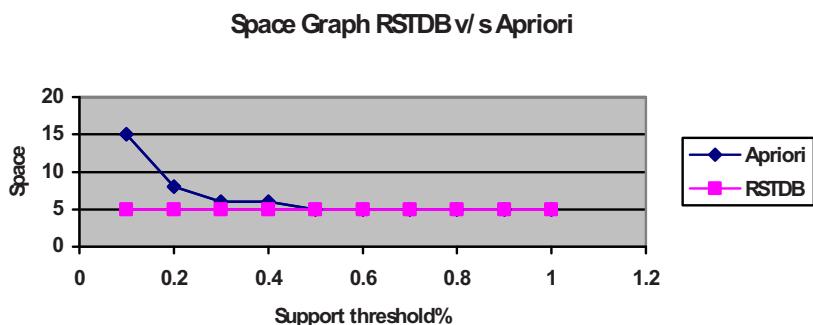
### 3.2.2 RSTBD V/S APRIORI

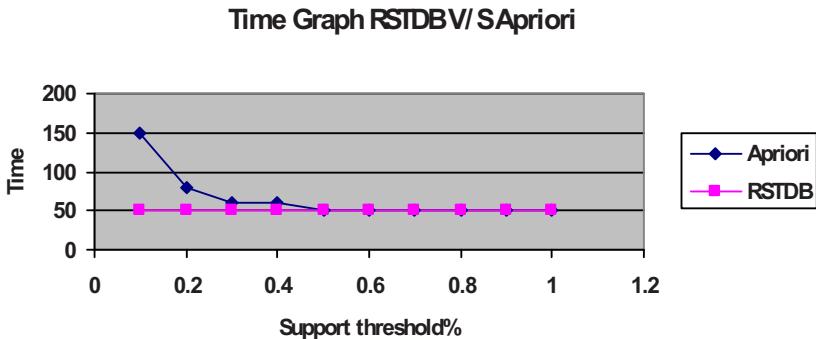
Consider the below database having five elements A, B, C, D, E. In the below table there are 10 transactions. We will mathematically show as to the difference between the two approaches:-

**Table 1.** TDB1 consisting of A, B, C, D, E

Transaction ID	Items
100	A
101	B
102	C
103	D
104	E
105	B
106	B
107	A
108	B
109	C

For the TDB the two approaches RSTDB and Apriori give the following results in terms of Space and execution time.

**Fig. 2.** Space complexity Graph for TDB1



**Fig. 3.** Time Complexity Graph for TDB1

Here, It is assumed that each scan require one unit time for the purpose and for space two units of data bytes is assumed to take space for each counter.

### 3.2.3 Limitations of RSTDB

1. It does not work for all conditions as it depends on the heuristic function
2. The increased efficiency is very less
3. Overhead is associated with heuristic function evaluation.

### 3.2.4 Experimental Result

1. The Proposed algorithm depends on the heuristic function.
2. It is more efficient for lower threshold values.
3. It depends both on the number of different items and total number of transactions.

## 4 Conclusion

Apriori is the simplest algorithm that is used for mining of frequent patterns from the transaction database. To reduce the number of scans required for the process of extraction of the frequent set the proposed algorithm reduces the number of scans by using both upward and downward closure property. Mining frequent patterns from large database plays an essential role in many data mining tasks and has broad applications. Most of the previous proposed methods adopt Apriori like candidate-generation-and-test approaches. However, those methods may encounter serious challenges when mining datasets with prolific patterns and or long patterns. RSTDB although not that efficient increases the efficiency in

frequent pattern mining by some amount. It mainly concerns with reducing the number of scans of database involved in mining process.

## Acknowledgement

At last, but not the least, we want to present our sincere thanks to HOD Computer Science & Engineering Prof. Y.K. Jain SATI Vidisha and also Prof. Dr. L.P.Pateria HOD MBA GGU Bilaspur for there help in our research work

## References

1. Agrawal, R., Imielinski, T., and Swami, A.N.: Mining association rules between sets of items in large databases. Proceedings of ACM SIGMOD International Conference on Management of Data, ACM Press, Washington DC, pp.207–216, May (1993)
2. Zaki, M.J.: Scalable Algorithms for Association Mining. IEEE Transactions on Knowledge and Data Engineering, vol.12, no. 3, pp. 372–390, May/June (2000)
3. Han, J., Pei, J., Yin, Y.:Mining Frequent Patterns without Candidate Generation. Proceedings of ACM SIGMOD International Conference on Management of Data, ACM Press, Dallas, Texas, pp. 1–12, May (2000)
4. Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., and Yang, D.:Hmine: Hyper-Structure Mining of Frequent Patterns in Large Databases. Proceedings of IEEE International Conference on Data Mining, pp. 441–448 (2001)
5. Pietracaprina, Zandolin, D.: Mining Frequent Item sets Using Patricia Tries. FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Item set Mining Implementations, Melbourne, Florida, Dec. (2003)
6. Grahne, G., Zhu, J.:Efficiently using prefix-trees in mining frequent itemsets. FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, Melbourne, Florida, December (2003)
7. Burdick, D., Calimlim, M., Flannick, J., Gehrke, J.:MAFIA: A Maximal Frequent Itemset Algorithm. IEEE Transactions on Knowledge and Data Engineering, 17, 1490–1505, Nov. (2005)
8. Zaki, M.J., Parthasarathy, S., Ogihara, M., Li, W.: New algorithms for fast discovery of association rules. Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, AAAI Press, pp. 283–286 (1997)
9. Shenoy, P., Haritsa, J.R., Sudarshan, S., Bhalotia, G., Bawa, M., Shah, D.:Turbo-charging vertical mining of large databases. Proceedings of ACM SIGMOD Intnatiional Conference on Management of Data, ACM Press, Dallas, Texas, pp. 22–23, May (2000)
10. Burdick, D., Calimlim, M., and Gehrke, J.: MAFIA: a maximal frequent item set algorithm for transactional databases. Proceedings of International Conference on Data Engineering, Heidelberg, Germany, pp. 443–452, April (2001)

11. Zaki, M.J., Gouda, K.:Fast vertical mining using diffsets. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington,D.C., ACM Press, New York, pp. 326–335, (2003)
12. Agrawal, R., Agarwal, C., Prasad, V.:A Tree Projection Algorithm for Generation of Frequent Item Sets. Parallel and Distributed Computing, pp. 350–371, (2000)
13. Singh, V.K., Shah, V., Jain Y.K., Shukla, A., Thoke, A.S., Singh, V.K., Dule, C., Parganiha, V.: Proposing an Efficient Method for Frequent Pattern Mining. has been Accepted for Oral Presentation at the Conference and publication in Proceeding of World Academy of Science, Engineering and Technology, Volume 36,International Conference on Computational and Statistical Sciences, Bangkok Dec 9 (2008)
14. Singh, V.K., Shah V.: Minimizing Space Time Complexity in Frequent Pattern Mining by Reducing Transaction Database Scanning and Using Pattern Growth Methods. To appear in Chhattisgarh Journal of Science and Technology (2008)
15. Singh, V.K., Shah V.: Minimizing Space Time Tradeoff in Frequent Pattern Mining Using Pattern growth Methods. Proceedings of Tech Acme 08 ,17–19 Oct Bhopal (2008)
16. Singh, V.K., Singh, V.K.: The Huge Potential of Information Technology. Proceedings of National Convention on Global Leadership: Strategies and Challenges for Indian Business, Feb 10–11, GGU Bilaspur (2007).

# **Privacy Preserving Data Mining: A New Methodology for Data Transformation**

A. K. Upadhayay<sup>1</sup>, Abhijat Agarwal<sup>2</sup>, Rachita Masand<sup>2</sup> and  
Rajeev Gupta<sup>3</sup>

<sup>1</sup>Amity School of Engineering and Technology, Noida, U.P., India  
aupadhayay@jpr.amity.edu

<sup>2</sup>Amity School of Engineering and Technology, Noida, U.P., India  
{abhijat.agarwal, rachita.masand}@gmail.com

<sup>3</sup>Rajasthan Technical University, Kota, Rajasthan, India  
rajeev\_eck@yahoo.com

**Abstract.** Today, privacy preservation is one of the greater concerns in data mining. While the research to develop different techniques for data preservation is on, a concrete solution is awaited. We address the privacy issue in data mining by a novel privacy preserving data mining technique. We develop and introduce a novel ICT (inverse cosine based transformation) method to preserve the data before subjecting it to clustering or any kind of analysis. A novel ‘privacy preserved k-clustering algorithm’ (PrivClust) is developed by embedding our ICT method into existing K-means clustering algorithm. This algorithm is explicitly designed with conversion to a privacy-preserving version in mind. The challenge was how to meet privacy requirements and guarantee valid clustering results as well. Simulation was carried out using Matlab. Our analysis and simulation show that this algorithm efficiently preserves the intended information on the one hand and yields valid cluster results on the other.

## **1 Introduction**

“Data mining is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. [5]” Though a misnomer, Data mining refers to art of mining new information in terms of patterns or rules, from databases of varying size and diverse nature. Data mining is connecting the three worlds of Databases, Artificial Intelligence and Statistics. The information age has enabled many organizations to gather large volumes of data. However, the usefulness of this data is negligible if “meaningful information” or “knowledge” cannot be extracted from it. Data mining, otherwise

known as knowledge discovery, attempts to answer this need. In contrast to standard statistical methods, data mining techniques search for interesting information without demanding *a priori* hypotheses.

Data mining is an emerging field and has already marked its strong debut. Apart from its historical use in identifying culprits of 9/11 through a mass of information and by the Central Intelligence Agencies both in US and Canada, Data mining is being primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to determine relationships among "internal" factors such as price, product positioning or staff skills, and "external" factors such as economic indicators, competition and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction and corporate profits. Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures. The National Basketball Association (NBA) is exploring a data mining application that can be used in conjunction with image recordings of basketball games to help coaches orchestrate plays and strategies. Data mining has been widely used in area of science and engineering, such as bioinformatics, genetics, medicine and electrical power engineering.

## 1.1 Privacy for Data Mining

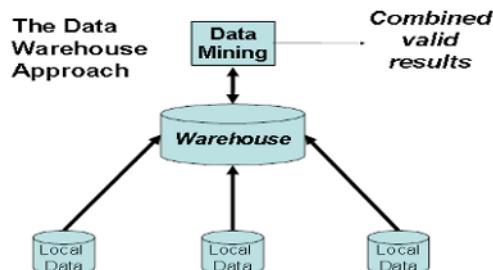
It happens quite often, during joint ventures, corporate collaboration, medical research etc. that highly sensitive information that could turn against its holder is to be leaked or is discovered. Information at times can be sensitive enough to be revealed to the discretion of others. Revealing sensitive information poses a threat to *privacy* of an organization or an individual holding that information.

- Consider a scenario in which two or more parties owning confidential databases wish to run a data mining algorithm on the union of their databases without revealing any unnecessary information. For example, consider NASA and ISRO, two separate space agencies that wish to conduct a joint research operation while preserving the privacy of their operations. In this scenario it is required to protect privileged information, but it is also required to enable its use for research or for other purposes. In particular, although the parties realize that combining their data has some mutual benefit, none of them is willing to reveal its database to any other party.

Generally when people talk of privacy, they say "keep information about me from being available to others" [2]. However, their real concern is that their information not be misused. The fear is that once information is released, it will be impossible to prevent misuse. Utilizing this distinction – ensuring that a data mining project won't enable further use or misuse of personal information – opens opportunities that *complete privacy* would prevent. To do this, we need technical and social solutions that ensure data will not be released.

### 1.1.1 Classification of

1. **Individual privacy [2]:** Typically people think of privacy as protecting individual data. ‘Personal data’ shall mean any information relating to an identified or identifiable natural person (data subject); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity; and specifies that data can be kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. The key element here is “identifiable”: As long as the data cannot be traced to an individual, his privacy is preserved. This is what the privacy preservation is all about, whether we take to the level of an individual or an enterprise.
2. **Corporate Privacy:** Another view is corporate privacy [2] – the release of information about a collection of data rather than an individual data item. I may not be concerned about someone knowing my birth date, mother’s maiden name, or social security number; but knowing all of them enables identity theft. This collected information problem scales to large, multi individual collections as well. A technique that guarantees no individual data is revealed may still release information describing the collection as a whole. Such “corporate information” is generally the goal of data mining, but some results may still lead to concerns (often termed a secrecy, rather than privacy, issue.) In the corporate model, the privacy restrictions need not be just on individually identifiable data. Often it is the body of data that must be protected. Protecting individual data items may not be enough we may need to protect against learning from the collection.



**Fig. 1.** The Data Warehouse Approach to mining distributed sources.

The typical approach to data mining of distributed data is to build a data warehouse containing all the data, and then mine the warehouse. This requires that the warehouse be trusted to maintain the privacy of all parties - since it knows the source of data, it learns site-specific information as well as global results.

### 1.1.2 Existing Techniques for Privacy Preservation

Here we present various existing techniques for privacy preservation such as data obscuration, randomization [4], data perturbation [1, 17], data transformation [9], cryptographic techniques [7, 11] etc. Other than these several techniques have been developed for privacy preservation in clustering over Vertically Partitioned Data [14]. We present the comparative advantages of the techniques developed till now in the following section.

#### Randomization

Randomization [1, 4] (encoding data with a random function/value/noise) of similar size and nature can be generated and applied to the data set. Randomization includes ADP (additive data perturbation) techniques [1, 18] as well as multiplicative data perturbation technique. A random value/noise when applied to the original data set produces a new encoded data set. The encoded data set contains exactly the same dimensions as of original data set and preserves its overall structure. This consequently minimizes the chances of invalid clustering results, thus preserving privacy of the data holder: an individual or an enterprise.

#### Data Obscuration

One approach to privacy is to obscure data: making private data available, but with enough noise added that exact values (or approximations sufficient to allow misuse) cannot be determined. Data obscuration is effective both in the web and corporate model. Obscuration can be done by the individual (if the receiver isn't trusted), or by the holder of data (to reduce concerns about breached security). Adding random noise to data values, then mining the distorted data lowers the accuracy of data mining results, but the research has shown that the loss of accuracy can be small relative to the loss of ability to estimate an individual item. We can reconstruct the original distribution of a collection of obscured numeric values, enabling better construction of decision trees. This would enable data collected from a web survey to be obscured at the source – the correct values would never leave the respondent's machine – ensuring that exact (susceptible to misuse) data doesn't exist.

Data obscuration techniques could also be used to ensure that otherwise identifiable data isn't individually identifiable. Re-identification experiments have shown that data that might be viewed as non-identifiable, such as birth date and postal code, can in combination allow identification of an individual. Obscuring the data could make re-identification impossible, thus meeting both the spirit and letter of privacy laws.

### Anonymity and Generalization

The goal during the privacy-preserving data mining process is to achieve results without revealing the identity of the individual users or any information that may result in identifying different people.

One method for protecting privacy is k-anonymity [12]. Roughly, the goal of k-anonymity is to only release data where for all possible queries, at least k results will be returned. To achieve this result, generalization (Table 1) and suppression techniques are used. In generalization techniques, some attributes are replaced with more general values so that k people will be found with any attribute value. For example, exact ages are replaced by some age ranges. In suppression techniques, data points that may cause too much generalization may be eliminated or a column that has identifying information can be deleted.

**Table 1.** Anonymous Table: Table with generalized values for attribute ‘Age’ and ‘Zip Code’

Row	Age	Sex	Zip Code	Disease
1	[1 , 10]	M	[10001 , 15000]	Gastric ulcer
2	[1 , 10]	M	[10001 , 15000]	dyspepsia
3	[1 , 10]	M	[15001 , 20000]	pneumonia
4	[1 , 10]	M	[15001 , 20000]	bronchitis
5	[11, 20]	M	[20001 , 25000]	flu
10	[21 , 60]	F	[30000 , 60000]	pneumonia

### Secure Multiparty Computation

The idea of Secure Multiparty Computation (SMC) is that the parties involved learn nothing but the results. Informally, imagine that we have a trusted third party to which all parties give their input. The trusted party computes the output and returns it to the parties. SMC enables this *without* the trusted third party. There may be considerable communication between the parties to get the final result, but the parties don’t learn anything from this communication. The computation is secure if given just one party’s input and output from those runs; we can *simulate* what would be seen by the party. In this case, to simulate means that the distribution of what is actually seen and the distribution of the simulated view

over many runs are computationally indistinguishable. We may not be able to exactly simulate every run, but over time we cannot tell the simulation from the real runs. Since we could simulate the runs from knowing only our input and output, it makes sense to say that we don't learn anything from the run other than the output. This seems like a strong guarantee of privacy, and has been used in privacy preserving data mining work. We must be careful when using Secure Multiparty Computation to define privacy. For example, suppose we use a SMC technique to build a decision tree from databases at two sites classifying people into high and low risk for a sensitive disease. Assume that the non-sensitive data is public, but the sensitive data (needed as training data to build the classifier) cannot be revealed. The SMC computation won't reveal the sensitive data, but the resulting classifier will enable all parties to estimate the value of the sensitive data. It isn't that the SMC was "broken", but that the result itself violates privacy.

## Two Party Computation

We now discuss issues specific to the case of two-party computation [11] where the inputs  $x$  and  $y$  are databases. Denote the two parties  $P_1$  and  $P_2$  and their respective private databases  $D_1$  and  $D_2$ . First, we assume that  $D_1$  and  $D_2$  have the same structure and that the attribute names are public. This is essential for carrying out any joint computation in this setting. There is a somewhat delicate issue when it comes to the names of the possible values for each attribute. On the one hand, universal names must clearly be agreed upon in order to compute any joint function. On the other hand, even the existence of a certain attribute value in a database can be sensitive information. This problem can be solved by a pre-processing phase in which random value names are assigned to the values such that they are consistent in both databases. Doing this efficiently is in itself a non-trivial problem. However, in our work we assume that the attribute-value names are also public (as would be after the above-described random mapping stage). Next, as we have discussed, each party should receive the output of some data mining algorithm on the union of their databases,  $D_1 \cup D_2$ . We do not assume any "trusted" third party who computes the joint output. The two party concept can be scaled to multiple party concept.

## 1.2 Clustering

Data mining involves clustering. Clustering (or unsupervised learning) is defined as classification of objects in different clusters based on *intra class similarity* and *inter class dissimilarity*. A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.

### 1.2.1 Types of Data Clustering Algorithms:

**Hierarchical Clustering Algorithm:** Hierarchical clustering [5] builds (agglomerative), or breaks up (divisive), a hierarchy of clusters. The traditional representation of this hierarchy is a tree (called a dendrogram), with individual elements at one end and a single cluster containing every element at the other. Agglomerative algorithms are also known as bottom-up, whereas divisive (top-down) algorithms begin at the root.

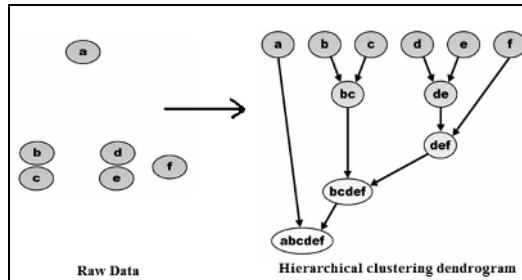


Fig. 2. Figure showing the concept of hierarchical clustering algorithm.

**K-medoids:** Here the initial representative objects are chosen arbitrarily. The iterative process of replacing representative objects with non representative objects continues as long as the quality of the resulting clustering is improved using a cost function. If the clustering is over binary objects, medoids need to be used. A medoid is just a bit string that minimizes the sum of distances to all objects in the cluster

**K-means Clustering:** The  $K$ -means clustering assigns each point to the cluster whose centre (also called centroid) is nearest. The centre is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. The goal is to divide the objects into  $K$  clusters such that some metric relative to the centroids of the clusters is minimized. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets.

## 2 Basic Concepts: A Review

In order to fully understand and appreciate the technique developed in this paper, reviewing of few basic concepts is indispensable.

## 2.1 Data Set

Every entity in this universe can be identified by various properties or characteristics it possesses. Let us say if we group together  $m$  of these similar types of objects or entities such that each object in the group has  $n$  number of attribute, we get data-set containing  $m$  records of  $n$  dimensions. Every data set is represented as a collection of entities or objects and each of these entities may have several attributes or dimensions. It is on this data set  $D$  of size  $[m \times n]$  clustering is performed.

$$D_{m \times n} = \begin{bmatrix} a_{11} & \cdots & a_{1k} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2k} & \cdots & a_{2n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mk} & \cdots & a_{mn} \end{bmatrix} \quad (1)$$

Sometimes it is important to normalize the data so that any inconsistencies, if present in the data set are removed and all the values fall within particular range, say between  $a$  and  $b$ . Two main techniques for normalization are min-max normalization and z-score normalization (aka. zero-mean normalization). The former is useful when min and max of an attribute are known whereas latter saves the day in case min and max are unknown or outliers are present[10].

## 2.2 The Heuristics for Driving Clustering

Several methods exists to group records in to the cluster so that inter cluster dissimilarity and intra cluster similarity is achieved. Three most popular methods employed to guide a clustering algorithm are ‘Euclidean distance’ (2) and ‘Cityblock or Manhattan distance’ (3) and standard Euclidean distance metric (4).

$$d(i, j) = \left[ \sum_{k=1}^n (a_{ik} - b_{jk})^2 \right]^{\frac{1}{2}} \quad (2)$$

$$d(i, j) = \sum_{k=1}^n |a_{ik} - b_{jk}| \quad (3)$$

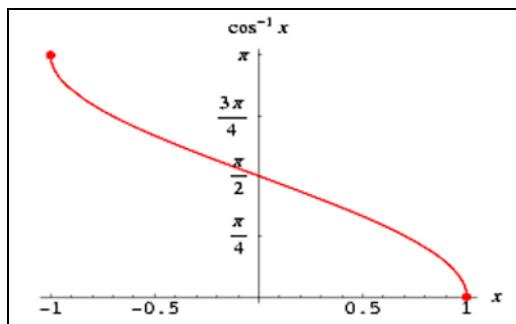
$$d(i, j) = \left[ \sum_{k=1}^n \frac{(a_{ik} - b_{jk})^2}{\sigma_{ij}} \right]^{\frac{1}{2}} \quad (4)$$

Here  $i=(a_{i1}, a_{i2}, \dots, a_{in})$  and  $j=(a_{j1}, a_{j2}, \dots, a_{jn})$  are  $n$ -dimensional data entities.

These formulae are used to find similarity or dissimilarity between objects of a data-set, which determines the end results of clustering. Thus they act as heuristics to guide clustering (Sect.1.2) algorithm (e.g. K-means) in determining valid placement of records under various clusters.

### 2.3 Inverse Trigonometric Functions

The inverse trigonometric functions [8] are the inverse functions of the trigonometric functions, written  $\cos^{-1}x$ ,  $\cot^{-1}x$ ,  $\csc^{-1}x$ ,  $\sec^{-1}x$ ,  $\sin^{-1}x$ , and  $\tan^{-1}x$ . Having known for their multiple valued characteristics these are the functions that assume two or more distinct values in their range for at least one point in their domain. Having known this fact we chose to analyze various inverse trigonometric functions and found an interesting thing about one of them. Let us analyze the domain of  $\cos^{-1}x$  which is  $x : -1 \leq x \leq 1$  as seen in Figure. 3.



**Fig. 3.** Graph depicting the domain of inverse cosine function.

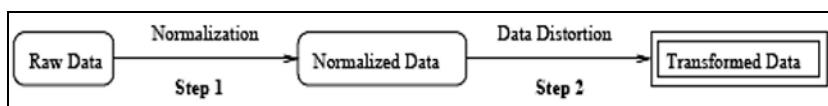
The interesting fact about  $\cos^{-1}x$  function is that if you use the function beyond its domain (i.e. if  $x > 1$ ) it yields a complex number whose real part is zero and only a *unique* imaginary value exists. By using this property of inverse cosine function we have developed a unique approach to encode the data and preserve its privacy.

### 3 Inverse Cosine-Based Transformation Method

In this section of our paper, we introduce our technique of Inverse cosine based transformation (ICT). This technique protects the actual values of attributes and preserves them before subjecting them to clustering. Our method also preserve the distances between the various attribute values since the encoding of each data item results in unique attribute value corresponding to its original value.

#### 3.1 The Process

We presume that the attribute values are normalized. The major steps to transform raw data into encoded data (privacy preserved data) are illustrated in Fig. 4.



**Fig. 4.** Major steps to data transformation

Step one deal with the process of normalization of raw data. Normalization process may be selected depending upon the nature of data as explained in Sect 2.1. Secondly the data is distorted. There may be several stages in the process of data distortion depending upon the method adopted.

#### 3.2 The Approach

After normalization we add a domain defying value (DDV) to each attribute value so that it would surpass the upper limit ( $x \leq 1$ ) in the domain of inverse cosine function. This would give us access to *imaginary values* generated by applying the transformation function. For example, we wish to encode '0.25'. After applying DDV we end up with 200. Now,  $\cos^{-1}(200)$  would yield  $0+5.9915i$ , a complex number. We may get rid of the imaginary 'i' (iota) by simply multiplying the complex number with another iota 'i'. Resultant value becomes -5.9915 which is fairly different from what we started off with. The algorithm is summarized below.

#### The ICT Algorithm

The process to encode the data attributes employs following essential steps:

**Input:**  $D_{(m \times n)}$ ,  $DDV$

**Output:**  $D^{T2}$

**Step 1.** Prepare the data by inducing **DDV (domain defying value)** to the original data set  $D_{(m \times n)}$  to yield  $D^{T1}$  such that all the attribute values in data set  $D^{T1}$  are greater than one.

$$D_{(m \times n)}^{T1} = DDV + D_{(m \times n)} \quad (5)$$

The value of DDV is obtained by using a **DDF** (domain defying function) which may easily be user defined and selected/formulated as per user's choice. After the normalization process it is the second step towards securing the data.

**Step 2.** Apply the encoding function  $E_{ICT}$  on  $D^{T1}$  to obtain  $D^{T2}$ .

$$D^{T2} = E_{ICT}(D^{T1}) \quad (6)$$

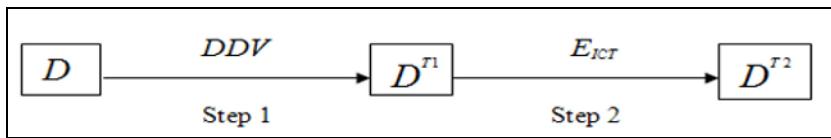
This encoding function transforms each attribute value of  $D^{T1}$  and results in new dataset  $D^{T2}$  which contains encoded values completely different from the original dataset thus marking the concluding step of ICT algorithm.

### The Encoding Function $E_{ICT}$

The encoding function takes dataset  $D_{(m \times n)}$  (all the attribute values in the dataset are greater than 1) as input and transforms each attribute value to new encoded value.

$$E_{ICT}(D_{(m \times n)}) = ICT(D_{(m \times n)}) \quad (7)$$

This transformation results in new encoded value for each corresponding attribute value by applying ICT (inverse cosine transformation function) which could be *customized* as per the requirements. It also maintains the original dissimilarity ratio among the data objects. This method preserves the original structure of the dataset which in turn produces valid clustering results when encoded data is subjected to clustering. Thus there is no privacy/accuracy trade-off [1]. The steps are summarized in Figure.5.



**Fig. 5.** Steps to encode data using ICT based approach.

### 3.3 The Algorithm: *PrivClust*

The *PrivClust* algorithm embeds our ICT based method for privacy preservation. It produces valid clustering results for the encoded dataset. The algorithm is listed below.

**Input:**  $D_{(m \times n)}$ , DDV, K.

**Output:** Privacy Preserved K clusters for  $D_{(m \times n)}$ .

1. Encode the data by ICT Method.
2. Randomly generate k clusters and determine the cluster centroids.
3. Assign each point to the nearest cluster centre using Euclidean Distance criteria. For two n-dimensional data points (records)  $r_j$  and  $r_k$ . the Euclidean distance between  $r_j$  and  $r_k$  is given by:

$$\text{Distance}(r_j, r_k) = \sqrt{|r_{j1} - r_{k1}|^2 + |r_{j2} - r_{k2}|^2 + \dots + |r_{jn} - r_{kn}|^2} \quad (8)$$

4. Re-compute the new cluster centroids mean) for each cluster.

$$\overline{C_i} = \left( \frac{1}{n} \sum_{\forall j \in i} r_{ji}, \dots, \frac{1}{n} \sum_{\forall j \in i} r_{jm} \right) \quad (9)$$

5. Repeat the two previous steps until convergence condition: The terminating condition is minimizing square error criterion (i.e. until there is no change in the assignment of cluster centroids). For clusters  $C_1, \dots, C_k$  and means  $m_1, \dots, m_k$ . the formula for ESS is.

$$\text{Error} = \sum_{i=1}^k \sum_{\forall r_j \in C_i} \text{Distance}(r_j, m_i)^2 \quad (10)$$

## 4 Performance of ICT Method

We bisect the performance into ‘accuracy’ and ‘security’. In this section first we show that the ICT method is highly accurate and extremely secure, with the help of an example. Though several methods for quantifying privacy [1, 16, 17] are available, we measure how closely an attacker may estimate the original data values by analyzing it from the attackers’ perspective. Further we illustrate and seal the performance of ICT method by embedding it with the k-means clustering and running this combined algorithm on ‘Iris’ data set. The results are shown for simple K-means clustering performed on original IRIS dataset without privacy preserving technique as well as the combined algorithm ***PrivClust*** which embeds our **ICT** method into K-means clustering algorithm.

### 4.1 ICT Method: Accuracy and Security

We prove accuracy of our method with the help of an example. To illustrate further we show that our method holds impeccable accuracy regardless of the size of dataset or attribute values. It can transform with equal ease a value with sixty decimal places and a value comprising seventy digits before decimal point. We lift a fragment of data from IRIS dataset and prove the accuracy of our method.

**Table 2** The Original Dataset

Sepal Length	Sepal Width	Petal Length
5.1000	3.5000	1.4000
4.9000	3.0000	1.4000
4.7000	3.2000	1.3000
4.6000	3.1000	1.5000

**Table 3.** The Normalized dataset

Sepal Length	Sepal Width	Petal Length
1.2402	1.3887	0
0.3382	-	0
	0.9258	-
	0.0000	-
0.5637		1.2247
-	-	1.2247
1.0147	0.4629	

**Table 4.** After applying DDV to Normalized data

Sepal Length	Sepal Width	Petal Length
20.459	14.043	4.200
0	8	0
16.106	6.6374	4.200
8		0
11.755	9.6000	2.485
1		4
9.5791	8.1187	5.914
		6

**Table 5.** The Encoded Dataset for Table 4.

Sep al Length h	Sep al Width	Pet al Length h
3.71	3.33	2.11
10	41	37
	2.58	2.11
3.4714	01	37
	3.15	1.56
56	22	04
2.95	2.78	2.46
00	35	33

On comparing the dissimilarity matrix of original as well as encoded data we find that they are exactly the same. Thus our method did not result in loss of accuracy of the original data. The distance between the data elements is maintained even after preserving the real data values.

The ICT method is completely secure. This may be understood by the fact that the method of encoding the original data does not limit it to a domain of values e.g. after normalization all the values in a given dataset lie within a specific range (say between -2 and 2) depending upon the normalization method applied. Our method is not dependent upon this condition. Thus the encoded values do not lie in a particular domain, which makes it nearly impossible for anyone to guess the methodology used to encode the data values. Even reverse engineering would yield futile results. Secondly, the user defined DDV (domain defying value, Sect. 3.2) when introduced to the dataset obscure the original data values. Finally, the data is encoded or distorted by our ICT method, thus sealing its security.

Attacker may even be tempted to normalize the encoded data in order to reverse the encoding process but this would lead to change in the dissimilarity matrix as observed from Table 6 and Table 7. This occurs due to the reason that variance of Normalized data is [1.000; 1000; 1.000] which is different from variance of transformed data [0.1132; 0.1022; 0.1393].

**Table 6.** The Dissimilarity matrix for Table 5.

0				
0.7912	0			
0.8720	0.7378	0		
1.0022	0.6599	0.9413	0	

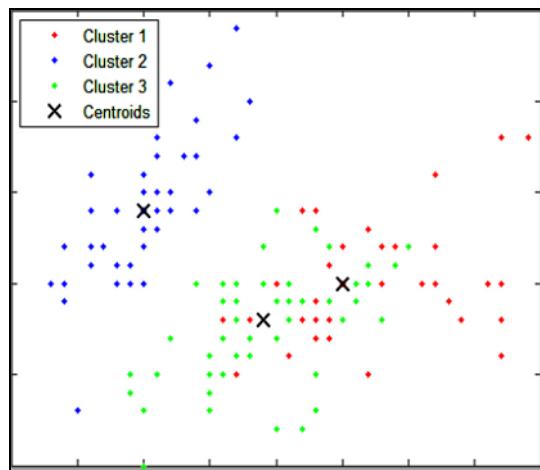
**Table 7.** Dissimilarity matrix corresponding to Table 5 after Normalization

0				
2.4642	0			
2.5198	2.1056	0		
2.9932	1.9192	2.5501	0	

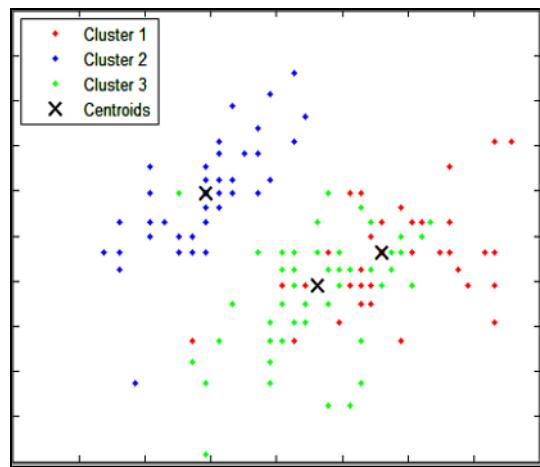
Now the only option remains is a brute force attack, in which an attacker tries all the possible methods or combinations to break the code. This may lead to pure frustration in the mind of an attacker and he would wisely wish to quit trying.

#### 4.2 Clustering Results for Iris Data

Here we have taken ‘Iris’ data and performed clustering using serial k-means (without privacy preservation) and then using *PrivClust* (our privacy preserved k-means clustering algorithm). Results observed from applying both the algorithms are shown and compared. Fig. 6 shows the clustering results for Iris data in its unaltered form and Fig. 7 shows the clustering results on encoded Iris data using ICT based *PrivClust* algorithm.



**Fig. 6.** Resulting clusters before privacy preserved K-means clustering using unaltered iris data and K-means clustering.



**Fig. 7.** Resulting clusters after privacy preserved K-means clustering using *PrivClust*

What we like to show here is that the over all structure of original dataset is maintained even after encoding. This will not lead to abrupt and invalid results from clustering. It is necessary for an organization to maintain its own privacy as

well as the corporate relations and contractual obligations. Our ICT based technique definitely assists in this process.

## 5 Applications of Our Work

1. In the world of collaborations, mergers, and joint business practices it is imperative for individual organization to see that its sensitive information does not go overboard. Major projects require multi entrepreneurs and multi financers, each one of which collects data and information which is pooled but only for the project. As each project entails a different consortium of entrepreneurs and financiers, no one would like its data to be used by the others for other projects. This may be difficult as each wants to preserve its privacy and still do business that too without straining mutual relationship.
2. Likewise for a medical institute, preserving the sensitive details (e.g. identity of HIV patients) is necessary during a *medical research* to avoid any level of discomfort or agony to its individuals.
3. Similarly in banking sector where financial details of account holders are kept under confidence, that may be of interest to competitors, other service providers as well as cheats, our privacy preservation technique would be of much help to the institution and would infuse *trust* in the consumers and ultimately result in business growth and *retention of high value customers* by preventing their poaching, which is necessary for survival in the cut throat competitive market. Further, banks and other financial service providers are converging for different financial services and making tie-ups with others like mutual funds, insurance companies, share depositories etc. Banks are also using common facilities like ATM providers, credit card companies, which would require privacy of each participant's sensitive data.
4. As the governments are getting increasingly computerised and are growingly interacting with private enterprises under public private partnership and bilateral/multilateral projects, *preservation of information of national interest* is vital for every participant government.
5. The crank call menace could also be managed with the help of our privacy preservation technique.

In short, our work on privacy preservation will be highly beneficial in the fields of business analysis, scientific and medical research, banking, collaborative defence practices, corporate joint ventures etc.

## 6 Conclusion

In this research we proposed a new technique called ICT (inverse cosine based transformation) for privacy preservation in data mining. We proved its accuracy and security in our analysis. This technique when embedded with K-means clustering algorithm resulted in a new algorithm '*PrivClust*'. ICT method can be embedded into any clustering algorithm to achieve privacy preservation and furthermore it may be even be used independently of any clustering algorithm for simply securing the data by encoding. The algorithm was simulated using Matlab and results were analyzed. As a result we have achieved privacy preserved K clusters for the Iris dataset using our *PrivClust* algorithm. Our method proved successful in achieving privacy preservation as well as producing valid clustering results. Finally the results show that there was no privacy/accuracy trade-off.

## 7 Future Work

Present holds the key to future and future becomes our present. So we would say that we are presently working on ideas to implement our algorithm on real data and data in the form of pictures, documents, audio, video etc. The purpose is to generalize the applicability of our method as much as possible and make it market ready. We would like to explore the possible customization options for generating DDV and defining  $E_{ICT}$ , which would diversely optimize the security of our method.

Our future work also includes widening the scope of our research to include classification, Artificial Neural Networks (ANNs), A.I. and fuzzy logic. We would also like to extensively study the brain so as to comprehend its functioning and demystify the human mechanism of privacy preservation and brain's ability to cluster and classify the mammoth of data which it receive and churns daily. It is immensely mind boggling to even think about how our brilliant brain works. The thought inspires us for the next project. We may express our humble belief in,

***“Neither the road ends here nor is it the destination,  
Keep walking my friend, the journey has just begun”***

\*

## References

1. Aggarwal, C.C., and Yu P.S.: Privacy Preserving data mining, Springer (2008)
2. Clifton C., Kantarcioglu M., Vaidya J.: Defining Privacy for Data Mining. Purdue University, West Lafayette.
3. Elmasri, N., Gupta S.: Fundamentals of Database Systems, Pearson Education, Inc, First Impression, (2006)

4. Evfimievski, A.: Randomization in Privacy-Preserving Data Mining. In SIGKDD Explorations, 4(2): 43-48, December (2002)
5. Hann, J., Kamber M.: Data Mining concepts and techniques, Elsevier, 2ed. (2006)
6. Jagannathan, G., Pillaipakkamnatt, K., Wright, R.N.: A New Privacy-Preserving Distributed k-Clustering Algorithm in proceedings of 2006 SIAM international conference on data mining on SDM-(2006)
7. Lindell, Y., Pinkas, B.: Privacy Preserving Data Mining, Advances in Cryptology -- Crypto '00 Proceedings, LNCS 1880, Springer-Verlag, pp. 20-24, August 2000. A full version appeared in the Journal of Cryptology, Volume 15 - Number 3, (2002)
8. Mathworld, <http://mathworld.wolfram.com/InverseTrigonometricFunctions.html>
9. Oliveira, S. R. M., Zaïane, O. R.: Privacy Preserving Clustering By Data Transformation. In Proceedings of the 18th Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, October (2003), pp.304-318.
10. Oliveira, S. R. M., Zaïane, O. R.: Achieving Privacy Preservation When Sharing Data for Clustering. In Proceedings of the International Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB (2004), Toronto, Canada, August, (2004)
11. Pinkas, B.: Cryptographic Techniques for Privacy-Preserving Data Mining SIGKDD Explorations, the newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining, January (2003)
12. Sweeny, L.: Achieving k-anonymity privacy protection using generalization and suppression. (2002) CMU.
13. Upadhyay, A.K., Gupta R., Kumar R.: Analytical model for revised K-clustering algorithm for privacy preservation in data mining. RACE (2007) at BEC Bikaner, IEEE sponsored international conference.
14. Vaidya, J., Clifton, C.: Privacy-Preserving K-Means Clustering over Vertically Partitioned Data. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August (2003) pp.206-215.
15. Wikipedia – The free encyclopedia, [www.wikipedia.org](http://www.wikipedia.org)
16. Agrawal, R., Srikant, R.: Privacy-Preserving Data Mining in proceedings of (2000) ACM SIGMOD Conference on Management of Data, pages 439-450, Dallas, TX, May 14-19 (2000). ACM.
17. Adam, N. R., Wortmann, J. C.: Security-Control Methods for Statistical Databases. ACM Computing Surveys, 21(4):515-556, Dec. (1989)
18. Murlidhar, K., Parsa, R., Sarathy, R.: A General Additive Data Perturbation Method for Database Security. Management Science, 45(10): 1399-1415, October (1999)

# Data Aggregation in Cluster based Wireless Sensor Networks

Shirshu Varma and Uma Shanker Tiwary

Department of Information Technology  
Indian Institute of Information Technology Allahabad, Allahabad, India  
[{shirshu, ust}@iiita.ac.in](mailto:{shirshu, ust}@iiita.ac.in)

**Abstract.** A wireless sensor network is a wireless network, in which we basically use tiny devices which monitor physical or environmental conditions at different regions. Thus the wireless sensor network basically comprises of sensing, communicating, computing, programming and control ingredients. We can design and configure sensor network applications designed that have better fault tolerance. In most applications sensor networks are deployed once and operate for a long period of time. In this respect managing and maintaining wireless sensor network becomes an important task. However, the problem of limited energy constraint presents a major challenge for the network deployment. Hence it clearly follows that in order to gather information from node to node we require data forwarding protocol. So the aim of any data forwarding protocol is to conserve energy to maximize the network lifetime. Sensor nodes are capable of performing in-network aggregation of data coming from more than one source. In this paper we have concentrated on energy consumption issue and designed energy efficient data aggregation protocol named E-BINA. Thus protocol using a cluster-based wireless sensor network is more relevant. Each cluster is executed independently and thus we obtain an energy efficiently data, which finally is aggregated in a cluster by this the lifetime of the cluster is also increased. We have used Network Simulator 2(NS-2) for simulating the protocol.

## 1 Introduction

A wireless sensor network is a wireless network consisting of tiny devices which monitor physical or environmental conditions such as temperature, pressure, motion or pollutants at different regions. The tiny device, known as sensor node, consists of a radio transceiver, microcontroller, power supply, and the actual sensor. Initially sensor network were used for military applications but now they are widely used for civilian application area including environment and habitat monitoring, healthcare application and so on. Normally sensor nodes are spatially

distributed throughout the region which has to be monitored; they self-organize in to a network through wireless communication, and collaborate with each other to accomplish the common task. With the going time, sensor nodes are becoming smaller, cheaper, and more powerful which enable us to deploy a large-scale sensor network.

Basic features of sensor networks are self-organizing capabilities, dynamic network topology, limited power, node failures & mobility of nodes, short-range broadcast communication and multi-hop routing, and large scale of deployment [12]. The strength of wireless sensor network lies in their flexibility and scalability. The capability of self-organize and wireless communication made them to be deployed in an ad-hoc fashion in remote or hazardous location without the need of any existing infrastructure. Through multi-hop communication a sensor node can communicate a far away node in the network. This allows the addition of sensor nodes in the network to expand the monitored area and hence proves its scalability & flexibility property. Wireless sensor network face several challenges such as physical resource constraints, ad-hoc deployment, fault-tolerance, scalability, and quality of service.

It is widely accepted that the energy consumed in one bit of data transfer can be used to perform a large number of arithmetic operations in the sensor processor [13]. Moreover in a densely deployed sensor network the physical environment would produce very similar data in close-by sensor nodes and transmitting such data is more or less redundant. Therefore, all these facts encourage using some kind of grouping of nodes such that data from sensor nodes of a group can be combined or compressed together in an intelligent way and transmit only compact data. This can not only reduce the global data to be transmitted and localized most traffic to within each individual group, but reduces the traffic and hence contention in a wireless sensor network. This process of grouping of sensor nodes in a densely deployed large-scale sensor network is known as clustering. The intelligent way to combined and compress the data belonging to a single cluster is known as data aggregation.

The main challenges with wireless sensor network are how to provide maximum lifetime to network and how to provide robustness to network. As sensor network totally rely on battery power, the main aim for maximizing lifetime of network is to conserve battery power or energy. In sensor network, the energy is mainly consumed for three purposes: data transmission, signal processing, and hardware operation. It is said in [4] that 70% of energy consumption is due to data transmission. So for maximizing the network lifetime, the process of data transmission should be optimized. The data transmission can be optimized by using efficient routing protocols and effective ways of data aggregation. Data aggregation protocols aims at eliminating redundant data transmission and thus improve the lifetime of energy constrained wireless sensor network. In wireless sensor network, data transmission took place in multi-hop fashion where each node forwards its data to the neighbor node which is nearer to sink. That neighbor node performs aggregation function and again forwards it on. But performing data forwarding and aggregation in this fashion from various sources to sink causes significant energy waste as each node in the network is

involved in operation. Since closely placed nodes may sense same data, above approach cannot be considered as energy efficient. An improvement over the above approach would be clustering where each node sends data to cluster-head (CH) and then cluster-head perform aggregation on the received raw data and then send it to sink. Performing aggregation function over cluster-head still causes significant energy wastage. In case of homogeneous sensor network cluster-head will soon die out and again re-clustering has to be done which again cause energy consumption. So here we have presented an algorithm that performs data aggregation within a cluster and thus reducing the load of aggregation at cluster-head to provide energy efficiency for maximizing network lifetime.

## 2 Related Works

Most of the work done till now on in-network aggregation mainly deals with problem of forwarding packets from source to sink, to facilitate aggregation therein. Actually the main idea behind were to enhance existing routing protocols such that they can efficiently aggregate data. So till now, most of the data aggregation techniques fall under three categories. They are tree-based approaches, multi-path approaches, and cluster-based approaches. There also some hybrid approaches that combines any of the three techniques above. The simplest way to aggregate data is to organize the nodes in a hierarchical manner and then select some nodes as the aggregation point or aggregators. The tree-based approach perform aggregation by constructing an aggregation tree, which could be a minimum spanning tree, rooted at sink and source nodes are considered as leaves. Each node has a parent node to forward its data. Some tree based approaches were presented in [14], [3], and [8].. The idea behind is that each node can send the data to its possibly multiple neighbors by exploiting the wireless medium characteristic.. Some multi-path approaches were presented in [15] and [5]. In cluster-based approach, whole network is divided in to several clusters and each cluster has a cluster-head which is selected among cluster members. W. Choi et al. in [1] present a two-phase clustering (TPC) scheme. Phase I of this scheme creates clusters with a cluster-head and each node within that cluster form a direct link with cluster-head. In phase II, each node within the cluster searches for a neighbor closer than cluster-head which is called data relay point and setup up a data relay link. H. Cam et al. in [4] present energy efficient and secure pattern based data aggregation protocol which is designed for clustered environment. This protocol says that instead of sending raw data to cluster-head, the cluster members send corresponding pattern codes to cluster-head for data aggregation.

## 3 E-BINA

We have designed our protocol which takes the merits of both cluster-based and tree-based approach. E-BINA assumes a cluster-based wireless sensor network

and applies tree-based approach inside each cluster.. The process of aggregation tree construction requires the sensor nodes to reduce their transmission power as sensor nodes now have to send their data packets to the neighbor node which is selected as parent. Energy consumed in wireless transmission is directly proportional to the square of the distance between nodes in communication [13]. Since cluster-member node now sends their data packets to the neighbor node instead of cluster-head, the transmission distance is reduced and hence the energy consumption of the sensor node. Likewise, overall energy consumption of sensor nodes in a cluster is reduced and so for the whole sensor network. Hence overall network lifetime will be increased.

Energy-aware Balanced In-Network Aggregation (E-BINA) protocol is energy-aware as it has taken the residual energy of sensor node in to consideration while constructing the aggregation tree. The protocol also balances the network load by selecting different parent for a node according to the energy level remain in the sensor node during the aggregation tree construction process. Each parent node performs aggregation of data packet that it receives from its child nodes and hence the protocol justifies the in-network aggregation concept.

### 3.1 System & Energy Model

Consider a homogeneous network of  $n$  sensor nodes and a base station or sink node distributed over a region. We assume that the location of the sensor nodes and the base station are fixed and known priori. Each sensor node produces some information as it monitors its vicinity. We assume that the whole network is divided in to several clusters; each cluster has a cluster-head (CH). The clustering and the selection of cluster-head (CH) can be done by using any existing protocol like LEACH, such that cluster-head (CH) is maximum  $k$ -hop away from any node in cluster. We also assume that after the formation of cluster the transmission power of all nodes is adjusted in such a way that they can perform single hop broadcast. The operation of sending a packet to all single-hop neighbors refers to Single hop broadcast [8].

Our energy model for the sensor nodes is based on the *first order radio model* described in [17]. A sensor consumes  $E_{elec} = 50nJ/bit$  to run the transmitter or receiver circuitry and  $E_{amp} = 100pJ/bit/m^2$  for the transmitter amplifier. Thus, the energy consumed by a sensor  $i$  in receiving a  $l$ -bit data packet is given by,

$$E_{Rxi} = E_{elec} \cdot l \quad (1)$$

while the energy consumed in transmitting a data packet to sensor  $j$  is given by,

$$E_{Txij} = E_{elec} \cdot l + E_{amp} \cdot d_{ij}^2 \cdot l \quad (2)$$

where  $d_{ij}$  is the distance between nodes  $i$  and  $j$ .

### 3.2 Protocol Description

In a cluster-based wireless sensor network, our algorithm is designed to provide energy-aware in-network data aggregation in a cluster. Each cluster uses this algorithm independently. In a cluster, the nodes can be categorized as: one cluster-head (CH) and other cluster member node. The algorithm works in two phases: Configuration packet flow and Data packet flow that are described below.

#### 1) Configuration Packet Flow

Initially cluster-head broadcast configuration packet to all its neighbors. Configuration packet contains the following fields:

Node Id → location of node that each node know in prior

Hop Distance → distance from cluster-head in terms of hop count (set zero for CH)

Residual Energy → current energy in node

Each node upon receiving the broadcast configuration packet that is originated from cluster-head adds the sender of the packet in the list of its possible parents with its node id, hop distance, residual energy. After this the node again broadcast the configuration packet to all its neighbors by updating node id to its own id, incrementing hop distance by one and its own residual energy. This process continues until all the sensor nodes in cluster receive configuration packet. All nodes that broadcast the configuration packet do so by predefined and common signal strength that is known to all the nodes.

**Define:**

$E^r[i]$ : residual energy of node i

$d_h[i]$ : distance from CH in terms of hop count of node i

$E^d_{ij}$ : difference of residual energy of two nodes i & j

$t_e$ : threshold of residual energy difference of two nodes i & j for balancing load

**nid**: id of a node

S: set of configuration packets received by node i

**ParentSelection(nid)**

1. select two nodes j & k such that  $d_h[j]$  &  $d_h[k]$  is minimal in S
2. if ( $d_h[j] < d_h[k]$  AND  $|E^d_{jk}| < t_e$ )
3.     then return Node id of node j
4.     else if ( $d_h[j] == d_h[k]$  AND  $|E^d_{jk}| < t_e$ )
5.         then return Node id of node with  $\max(E^r[j], E^r[k])$
6.     endif
7.     else return Node id of node k
8. endif

**Fig. 1.** Procedure for the parent selection in the cluster

## 2) Data Packet Flow

When all nodes receives configuration packets, each node now select the parent to which it can forward the data packet. This basic parent selection procedure is shown in Fig. 1. Each node looks in to the list of all its possible parents. The neighbor node which has least hop distance, i.e. closest to cluster-head, is selected as parent by a node. In case when two neighbor nodes have the least but equal hop distance, the node checks the residual energy of two neighbor nodes. The neighbor node that has greater residual energy is now selected as parent. In both the cases, node also calculate the difference of residual energy of two neighbor nodes, which have least hop distance, and checks whether this difference is less than the threshold or not. If it is then the node selects the parent as usual. But if it is not then the node selects other neighbor node as its parent. This allows a sensor node that has more available resources to be selected as a parent node. This also balances the consumption of energy of nodes in the cluster and leads to die out of nodes nearly at same time.

After selecting the parent node, each node now forwards its data to its parent. When a parent node receives multiple data packets from its neighbor nodes, it performs aggregation operation by eliminating redundancy in the data. Each parent node checks the equation below:

$$| V_{Ni} - V_{Nj} | < K \quad (3)$$

where,  $V_{Ni}$  → data value of node i  
 $V_{Nj}$  → data value of node j  
 $K$  → redundancy factor

If this equation satisfies, the parent node perform aggregation by applying any aggregation functions like MIN, MAX, and AVG on the values of data packet and send only one packet while discarding other packets. But if this equation do not satisfies, the parent performs aggregation by simply concatenating two data packet in to one keeping value of both packets intact.

The selection of value for redundancy factor ( $K$ ) has a trade-off between precision and energy consumption. If the application wants more precision, it should select a low value for redundancy factor otherwise a high value. Selecting high value for  $K$  means sending only one value thus less number of bits needs to be transmitted and hence low energy consumption.

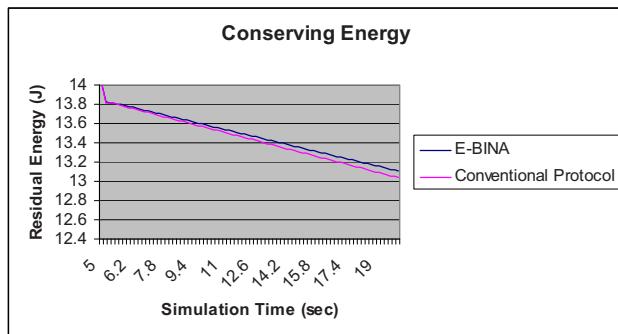
## 4 Simulation Analysis

We have chosen Network Simualtor-2 (NS-2) [20], in particular NS-2.29.3, as our tool to simulate the proposed protocol. A square field of 160m X 160m is taken

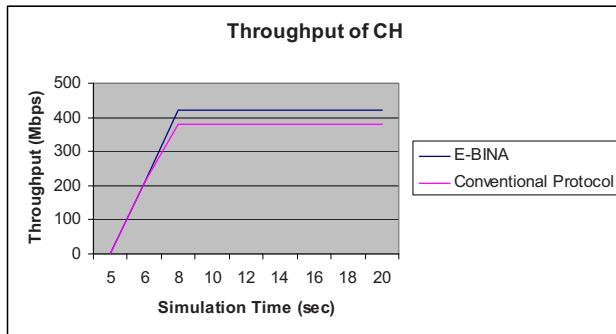
where 11 nodes are randomly deployed. One node is designated as cluster-head (CH) and one node is designated data source. The cluster-head is formed by the sink. Source node randomly sends packages with constant bit rate (CBR) to the sink. Packet size is 64 bytes, package rate is 5 pkt/s and each sensor node has a radio range of 40m. We choose  $14J$  as the sensor initial energy value,  $0.66W$  as the transmit power,  $0.395W$  as the receive power, and  $0.035W$  as the idle power. We assume that a sensor node consumes no energy when in sleep mode.

#### A) Conserving Energy

We first determine residual energy of the source node, which is defined as the remaining energy of a node and considered that as the metric to prove energy efficiency of our proposed protocol. Fig. 2 shows the significant reduction in energy consumption by using E-BINA when compared with conventional protocol.



**Fig. 2.** Residual energy of source as a function of time



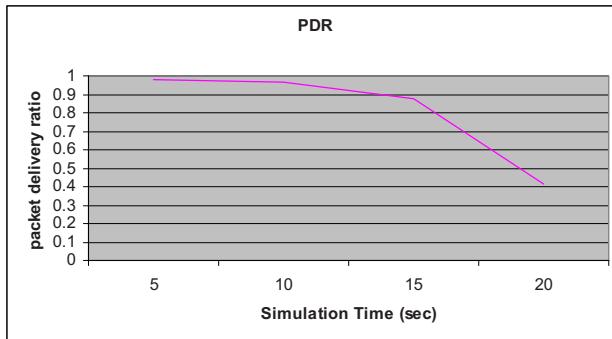
**Fig. 3.** Comparative throughput as a function of time

#### B) Throughput

We have also measured the throughput of the receiving node i.e. cluster-head node 10 in our scenario for both the cases. Throughput of a node is defined as the average rate of successful message delivery over a communication channel. Fig. 3 show that E-BINA achieves high throughput in comparison with conventional protocol.

#### C) Packet Delivery Ratio

Besides examining the network lifetime extension roughly via energy saving, we also evaluate the network efficiency influenced by E-BINA. Here, we measure the performance in term of data delivery ratio, which is defined as the number of received packets divided by the number of sent packets for a certain time period. Fig. 4 shows our simulation results and we find that this ratio does not change much while the network is alive and the stable performance of our protocol. When the energy of network is running out, the data delivery ratio collapses rapidly. This trend probably can be taken as a sign of the network death.



**Fig. 4.** Packet delivery ratio of the network in E-BINA

## 4 Conclusion

In this paper, we have proposed a data aggregation protocol E-BINA, which executes independently on each cluster in a cluster-based wireless sensor network and avoids aggregation only at cluster-head by constructing an aggregation tree rooted at cluster-head. The simulation result shows that when the data from source node is send to cluster-head through neighbors nodes in a multi-hop fashion by reducing transmission and receiving power, the energy consumption is low as compared to that of sending data directly to cluster-head. Future work will focus on the implementation of E-BINA in NS-2 as a separate module so that it could be tested more accurately. Also the effect of redundancy factor on energy consumption and overall performance of our protocol will be measured.

## References

1. Choi W., Shah P., Das S. :A Framework for Energy-Saving Data Gathering Using Two-Phase Clustering in Wireless Sensor Networks. in Proceedings of the International Conference on Mobile and Ubiquitous Systems: Networking and Services (MobiQuitous), Boston, 2004, pp. 203–212 (2004)
2. Lee M., Lee S. :Data Dissemination for Wireless Sensor Networks. in Proceedings of the 10<sup>th</sup> IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC'07), (2007)
3. Intanagonwiwat C., Govindan R., Estrin D., Heidemann J., Silva F. : Directed Diffusion for Wireless Sensor Networking. IEEE/ACM Transactions on Networking, Vol. 11, no. 1,(2003)
4. Cam H., Ozdemir S., Nair P., Muthuavinashiappan D. : ESPDA: Energy-Efficient and Secure Pattern-based Data Aggregation for Wireless Sensor Networks. In: Proceedings

- of IEEE Sensor- The Second IEEE Conference on Sensors, Toronto, Canada, Oct. 22-24pp. pp. 732-736 (2003)
- 5. Gatani L., Lo Re G., Ortolani M. :Robust and Efficient Data Gathering for Wireless Sensor Networks. in Proceeding of the 39th Hawaii International Conference on System Sciences (2006)
  - 6. Dasgupta K., Kalpakis K., Namjoshi P. : An Efficient Clustering-based Heuristic for Data Gathering and Aggregation in Sensor Networks. In: Proc. in IEEE WCNC, March, (2003)
  - 7. Fasolo E., Rossi M., Widmer J., Zorzi M. : In-Network Aggregation Techniques for Wireless Sensor Networks: A Survey. IEEE Wireless communication (2007)
  - 8. Lee M., Wong V.W.S.: An Energy-aware Spanning Tree Algorithm for Data Aggregation in Wireless Sensor Networks. IEEE PacRim 2005, Victoria, BC, Canada, (2005)
  - 9. Ding M., Cheng X., Xue G. : Aggregation tree construction in sensor networks. In Proc. of IEEE VTC'03, Vol. 4, Orlando, FL (2003)
  - 10. Romer K., Mattern F. : The Design Space of Wireless Sensor Networks, IEEE Wireless Communications, pp. 54–61, ( 2004)
  - 11. [http://en.wikipedia.org/wiki/Wireless\\_Sensor\\_Networks](http://en.wikipedia.org/wiki/Wireless_Sensor_Networks)
  - 12. Haenggi M. :Opportunities and Challenges in Wireless Sensor Networks in "Handbook of Sensor Networks: Compact Wireless and Wired Sensing Systems. Edited by M. Ilyas and I. Mahgoub, CRC Press ( 2004)
  - 13. Cordeiro C. D. M., Agrawal D. P. : Ad-hoc & Sensor Networks. World scientific publisher (2006)
  - 14. Madden S. et al. :TAG: a Tiny Aggregation Service for Ad-hoc Sensor Networks. In Proc. In OSDI 2002, Boston, MA, (2002)
  - 15. Nath S. et al., :Synopsis Diffusion for Robust Aggregation in Sensor Networks. In Proc. in ACM SenSys 2004, Baltimore, (2004)
  - 16. Kalpakis K., Dasgupta K., Namjoshi P. : Maximum Lifetime Data Gathering and Aggregation in Wireless Sensor Networks. In: Proceedings of IEEE Networks'02 Conference, (2002)
  - 17. Heinzelman W., Chandrakasan A.P., Balakrishnan H. : Energy-Efficient Communication Protocol for Wireless Microsensor Networks. In Proc. Of Hawaiian International Conference on System Science (2000)
  - 18. NRL's Sensor Network Extension to ns-2. <http://nrlsensorim.pf.itd.nrl.navy.mil/>
  - 19. Curren, D.: A survey of Simulation in Sensor Networks. [www.cs.binghamton.edu/~kang/teaching/cs580s/david.pdf](http://www.cs.binghamton.edu/~kang/teaching/cs580s/david.pdf), (2007)
  - 20. The Network Simulator NS-2, <http://www.isi.edu/nsnam/ns/>, January (2008)