

Journal of Advances in Information Technology

ISSN 1798-2340

Volume 1, Number 1, February 2010

Contents

EDITORIAL

Welcome Message from the Editor-in-Chief <i>A.C.M. Fong</i>	1
Introducing the Associate Editor-in-Chiefs <i>A.C.M. Fong</i>	2
Introduction to the Inaugural Issue <i>A.C.M. Fong</i>	3

REGULAR PAPERS

A Review of Machine Learning Algorithms for Text-Documents Classification <i>Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, and Khairullah Khan</i>	4
Multilingual Context Ontology Rule Enhanced Focused Web Crawler <i>Mukesh Kumar and Renu Vig</i>	21
Integrated Performance and Visualization Enhancements of OLAP Using Growing Self Organizing Neural Networks <i>Muhammad Usman, Sohail Asghar, and Simon Fong</i>	26
Design and Implementation of an Online Social Network with Face Recognition <i>Ray K.C. Lai, Jack C.K. Tang, Angus K.Y. Wong, and Philip I.S. Lei</i>	38
Dynamic Differential Evolution for Constrained Real-Parameter Optimization <i>Youyun Ao and Hongqin Chi</i>	43
Fuzzy Logic Based Position-Sensorless Speed Control of Multi Level Inverter Fed PMBLDC Drive <i>T.V. Narmadha and T. Thyagarajan</i>	52
On Performance of Multicast Delivery with Fixed WiMAX Telemedicine Networks Using Single-Carrier Modulation <i>Bernard Fong and Guan Yue Hong</i>	59

Welcome Message from the Editor-in-Chief

Dear Reader,

It is with much joy and anticipation that we celebrate the launch of Journal of Advances in Information Technology (JAIT) with this inaugural issue. On behalf of the JAIT Editorial Team, I would like to extend a very warm welcome to the readership of JAIT. I take this opportunity to thank our authors, editors and anonymous reviewers, all of whom have volunteered to contribute to the success of the journal. I am also grateful to Dr. George Sun and the staff at Academy Publisher for making JAIT a reality.

JAIT is dedicated to the rapid dissemination of high quality research papers on how advances in IT can help us meet the challenges of the 21st century, and to capitalize on the promises ahead. We welcome contributions that can demonstrate near-term practical usefulness, particularly contributions that take a multidisciplinary / convergent approach because many real world problems are complex in nature.

Barely a decade into the new millennium, we have witnessed significant events such as 9/11, SARS, and the South Asian Tsunami, to name but a few. There are also the on-going issues, such as demographic changes (population aging, internal migration and rapid urbanization), management of resources, and environment issues like the current debate about climate change. As IT practitioners and researchers, we aim to seek ways to harness the power of technology to meet some of these real world challenges, and to provide substance for making informed judgments on important matters. For example, modeling and prediction can help manage graying populations, model the spread of diseases or identify/mitigate future security threats; analysis of satellite imagery coupled with knowledge discovery can help manage deforestation, resource exploration or prepare future evacuation plans in response to natural disasters; large scale information processing can aid understanding of how (and to what extent) human activities can impact the environment and climate. The list of possibilities goes on.

As for the promises that lay ahead, IT has already become an integral part of everyday life. From commerce and government to scientific discovery, healthcare, education and entertainment, IT is indispensable and will continue to fuel further advances in all facets of human endeavors. With “e-everything”, we already see that IT has revolutionized the way we conduct business and learn. Further advances in IT will continue to change the way we live and play. As IT practitioners and researchers, we are responsible for making all this happen, and to some extent help shape the future. With technological advances also come social and legal changes. For example, popularization of the internet has led to a proliferation of virtual online communities, which are often formed on an ad hoc basis and typically characterized by a share of some common interests. Compared to a generation ago, people today interact frequently online and these changes inevitably have social implications.

JAIT provides an ideal forum for exchange of information on all of the above topics and more, in various formats: full length and letter length research papers, survey papers, work-in-progress reports on promising developments, case studies / best practice articles written by industry experts, and tutorials on up-and-coming technological breakthroughs. JAIT is published four times a year. To ensure rapid dissemination of information, we aim at completing the review process of each paper within 3 months of initial submission. We also publish special issues on relevant themes proposed by guest editors.

Finally, we wish to encourage more contributions from the scientific community and industry practitioners to ensure a continued success of the journal. Authors, reviewers and guest editors are always welcome. We also welcome comments and suggestions that could improve the quality of the journal.

Thank you. We hope you will find JAIT informative.

A.C.M. Fong
Editor-in-Chief
February 2010

Introducing the Associate Editor-in-Chiefs

I am pleased to introduce the distinguished Associate Editor-in-Chiefs of the JAIT.

A.C.M. Fong
Editor-in-Chief
 February 2010



Prof. Dr. Jinan Fiaidhi. Jinan Fiaidhi is Full Professor with tenure of Computer Science at the Lakehead University. She is also an Adjunct Research Professor with the University of Western Ontario. She received here graduate degrees in Computer Science from Essex University (PgD 1983) and Brunel University (PhD, 1986). During the period (1986-2001), Dr. Fiaidhi served at many academic positions (e.g. University of Technology (Asso. Prof and Chairperson), Philadelphia University (Asso. Prof), Applied Science University (Professor), Sultan Qaboos University (Asso. Prof.). Since late 2001, Dr. Fiaidhi is full Professor and Graduate Coordinator of Computer Science at Lakehead University, Ontario, Canada. Dr. Fiaidhi research is focused on mobile and ubiquitous learning utilizing the emerging technologies (e.g. Cloud Computing, Enterprise Mashups, Semantic Web). Dr. Fiaidhi research is supported by the major research granting associations in Canada (e.g. NSERC, CFI).

Moreover, Dr. Fiaidhi is a Professional Software Engineer of Ontario (PEng), Senior Member of IEEE, member of the British Computer Society (MBCS) and member of the Canadian Information Society (CIPS) holding the designate of ISP. Dr. Fiaidhi has intensive editorial experience (e.g. Editor of IEEE IT-Pro, Associate EiC of the Journal of Emerging Technologies in Web Intelligence).



Kin-Choong Yow received the BE degree with 1st class honours in electrical engineering from National University of Singapore, in 1993 and his PhD degree from Cambridge University, UK in 1998. He is currently an Associate Professor of Computer Engineering in the College of Engineering, Nanyang Technological University (NTU), Singapore. His research interests include Computer Vision, Wireless Communications and Computational Intelligence. He has more than 65 publications in international journals and conference proceedings, and he has served as reviewer for a number of premier journals and conferences, including the IEEE Wireless Communications and the IEEE Transactions on Education. He has been invited to give presentations at various scientific meetings and workshops, such as the CNET Networks Event 2002 as well as the Microsoft Windows Server 2003 Launch 2003. His pioneering work in Mobile and Interactive Learning won the HP Philanthropy grant in 2003 for applying Mobile Technologies in a Learning Environment. Also, in 2003, he was one of the only 2 Singaporeans to be awarded participation to the ASEAN Technology Program on Multi Robot Cooperation

Development held in KAIST, Korea. He was the winner of the NTU Excellence in Teaching Award 2005, and he won the Most Popular SCE Year 1 lecturer for 4 consecutive years from 2004-2007. He has led numerous student teams to National and International victories such as the IEEE Computer Society International Design Competition (CSIDC) 2001, the Microsoft Imagine Cup 2002, 2003 and 2005, and the Wireless Challenge 2003. He is also a member of the IEEE, ACM, and the Singapore Computer Society (SCS).

Introduction to the Inaugural Issue

In this inaugural issue, we present a variety of papers to cover, as much as possible, the breadth and depth of the intended scope of JAiT. We begin with a survey paper by Khan *et al.* on machine learning techniques for text document classification. With a proliferation of electronic documents on the web and elsewhere, it is increasingly important to be able to classify such e-documents for proper management. Their paper presents a timely review on some of the more prominent theories and methods of document classification and text mining for e-documents.

The second paper by Kumar and Vig presents a focused crawler that is enhanced by ontological rules. So-called focused crawlers have been developed primarily to seek and process relatively “untouched” web contents, such as some non-English documents that are not indexed by mainstream crawlers. The proposed focused crawler is intended for multilingual applications and the authors have applied their proposed crawler to mixed English and Hindi contents.

Online Analytical Processing (OLAP) is an important approach for mining multidimensional data. It has found widespread business applications, for instance, in decision support. In their paper entitled “Integrated Performance and Visualization Enhancements of OLAP Using Growing Self Organizing Neural Networks”, the authors present a novel architecture that reportedly can offer significant improvements over previous methods.

Online social networks have revolutionized the way people interact. In the next paper, Lai *et al.* present a social network that is enhanced with a face recognition and tagging feature. They also discuss the issues involved in designing and implementing such a system. Their paper therefore lays the groundwork for further developments in the drive towards enhancing the users’ experience of such online networks.

In the paper that follows, the authors introduce a dynamic differential evolution (D-DE) algorithm to solve constrained optimization problems. Three major improvements over the prior art have been reported, and the authors have performed experiments using six benchmark functions to substantiate their claim.

The last two papers underscore the wider applications of computing and information technology to the industry. The paper authored by Narmadha and Thyagarajan presents a multi level inverter fed Permanent Magnet Brushless DC Motor (PMBLDCM) with a simplified voltage control technique based on fuzzy logic. Sensing is through “indirect position sensing,” which is justified by the observation that position sensing came indirectly from voltage and current waveforms. Experiments and simulations conducted by the authors demonstrate the advantages of their approach.

In the final paper, the authors present an analysis of a fixed Worldwide Interoperability for Microwave Access (WiMAX) system that has the potential to provide a low cost solution for integrated multimedia access networks with a wide bandwidth. This makes it particularly suitable for telemedicine applications. They have presented results on comparing the distribution of video data using two modulation schemes, and have estimated the bandwidth utilization for continuous data transmission in remote patient monitoring applications.

We hope you will enjoy reading the papers published in this inaugural issue and find its contents to be very valuable.

A.C.M. Fong
Editor-in-Chief
February 2010

A Review of Machine Learning Algorithms for Text-Documents Classification

Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee*, Khairullah Khan

Department of Computer and Information Science,
Universiti Teknologi PETRONAS, Tronoh, Malaysia.

*Faculty of Science, Engineering and Technology,
Universiti Tunku Abdul Rahman, Perak Campus, Kampar, Malaysia.

(E-mail: aurangzebb_khan@yahoo.com, baharbh@petronas.com.my, leelh@utar.edu.my, khairullah_k@yahoo.com)

Abstract— With the increasing availability of electronic documents and the rapid growth of the World Wide Web, the task of automatic categorization of documents became the key method for organizing the information and knowledge discovery. Proper classification of e-documents, online news, blogs, e-mails and digital libraries need text mining, machine learning and natural language processing techniques to get meaningful knowledge. The aim of this paper is to highlight the important techniques and methodologies that are employed in text documents classification, while at the same time making awareness of some of the interesting challenges that remain to be solved, focused mainly on text representation and machine learning techniques. This paper provides a review of the theory and methods of document classification and text mining, focusing on the existing literature.

Index Terms— Text mining, Web mining, Documents classification, Information retrieval.

I. INTRODUCTION

The text mining studies are gaining more importance recently because of the availability of the increasing number of the electronic documents from a variety of sources. The resources of unstructured and semi structured information include the word wide web, governmental electronic repositories, news articles, biological databases, chat rooms, digital libraries, online forums, electronic mail and blog repositories. Therefore, proper classification and knowledge discovery from these resources is an important area for research.

Natural Language Processing (NLP), Data Mining, and Machine Learning techniques work together to automatically classify and discover patterns from the electronic documents. The main goal of text mining is to enable users to extract information from textual resources and deals with the operations like, retrieval, classification

(supervised, unsupervised and semi supervised) and summarization. However how these documented can be properly annotated, presented and classified. So it consists of several challenges, like proper annotation to the documents, appropriate document representation, dimensionality reduction to handle algorithmic issues [1], and an appropriate classifier function to obtain good generalization and avoid over-fitting. Extraction, Integration and classification of electronic documents from different sources and knowledge discovery from these documents are important for the research communities.

Today the web is the main source for the text documents, the amount of textual data available to us is consistently increasing, and approximately 80% of the information of an organization is stored in unstructured textual format [2], in the form of reports, email, views and news etc. The [3] shows that approximately 90% of the world's data is held in unstructured formats, so Information intensive business processes demand that we transcend from simple document retrieval to knowledge discovery. The need of automatically retrieval of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully apparent [4].

Market trend based on the content of the online news articles, sentiments, and events is an emerging topic for research in data mining and text mining community [5]. For these purpose state-of-the-art approaches to text classifications are presented in [6], in which three problems were discussed: documents representation, classifier construction and classifier evaluation. So constructing a data structure that can represent the documents, and constructing a classifier that can be used to predicate the class label of a document with high accuracy, are the key points in text classification.

One of the purposes of research is to review the available and known work, so an attempt is made to collect what's known about the documents classification and representation. This paper covers the overview of syntactic and semantic matters, domain ontology, tokenization concern and focused on the different machine learning techniques for text classification using the existing literature. The motivated perspective of the related research areas of text mining are:

Information Extraction (IE) methods is aim to extract specific information from text documents. This is the first

Aurangzeb Khan and Khairullah Khan are PhD Students, Department of Computer and Information Science at Universiti Teknologi PETRONAS, Tronoh, Malaysia.

Baharum Baharudin is an Assistant Professor at the Department of Computer and Information Science at Universiti Teknologi PETRONAS, Tronoh, Malaysia.

Lam Hong Lee is an Assistant Professor at the Faculty of Science, Engineering and Technology of Universiti Tunku Abdul Rahman, Perak Campus, located in Kampar, Malaysia.

(E-mail: aurangzebb_khan@yahoo.com, baharbh@petronas.com.my, leelh@utar.edu.my), Khairullah_k@yahoo.com)

Manuscript received May 28, 2009; revised September 7, 2009.

approach assumes that text mining essentially corresponds to information extraction.

Information Retrieval (IR) is the finding of documents which contain answers to questions. In order to achieve this goal statistical measures and methods are used for automatic processing of text data and comparison to the given question. Information retrieval in the broader sense deals with the entire range of information processing, from data retrieval to knowledge retrieval [7].

Natural Language Processing (NLP) is to achieve a better understanding of natural language by use of computers and represent the documents semantically to improve the classification and informational retrieval process. Semantic analysis is the process of linguistically parsing sentences and paragraphs into key concepts, verbs and proper nouns. Using statistics-backed technology, these words are then compared to the taxonomy.

Ontology is the explicit and abstract model representation of already defined finite sets of terms and concepts, involved in knowledge management, knowledge engineering, and intelligent information integration [23].

In this paper we have used system literature review process and followed standard steps for searching, screening, data-extraction, and reporting.

First of all we tried to search for relevant papers, presentations, research reports and policy documents that were broadly concerned with documents classification or text mining. We identified appropriate electronic databases and websites. Potentially relevant papers were identified using the electronic databases and websites, Such as IEEE Explore, Springer Linker, Science Direct, ACM Portal and Googol Search Engine. For best and consistent search a systematic search strategy was adopted. Proper keywords, queries, and phrases were derived from the desired research question. These keywords were arranged into categories and related keywords were arranged. Some facilities of digital libraries like sort by year etc were also used. The search keywords were refined to include only those words which have produced successful results. We used boolean logic for efficient searching, for example (Classification OR text OR recommendations). We also tried combination of words like Text Mining, Trend and Ontology analysis, Documents classification and Subjectivity Analysis etc.

Each search results were checked and assessed on screen to find relevance for inclusion and exclusion with the criteria that we made two categories of papers i.e. in or before 2000 and after 2000. The following studies were included: The result statements written in English, The research is conducted after 1980, Published and/or unpublished research, focused on documents classification, Machine Learning and Natural Language Processing (NLP). The non English writing and study before 1980 were excluded.

To find evidence and check the quality of papers we carried out an in-depth study of the results provided from the research. In our future work we will try to make this step

more strong and effective. We have tried to get some reports drawn using tables and graphs on the basis of existing studies.

The rest of the paper is organized as follows. In Section 2 an overview of documents representation approaches, Section 3 presents document classification models, in Section 4 new and hybrid techniques were presented. Section 5 consists of comparative study of different methods and finally in Section 6, some discussions and conclusion were made.

II DOCUMENTS REPRESENTATION

The documents representation is one of the pre-processing technique that is used to reduce the complexity of the documents and make them easier to handle, the document have to be transformed from the full text version to a document vector. Text representation is the important aspect in documents classification, denotes the mapping of a documents into a compact form of its contents. A text document is typically represented as a vector of term weights (word features) from a set of terms (dictionary), where each term occurs at least once in a certain minimum number of document. A major characteristic of the text classification problem is the extremely high dimensionality of text data. The number of potential features often exceeds the number of training documents. A definition of a document is that it is made of a joint membership of terms which have various patterns of occurrence. Text classification is an important component in many informational management tasks, however with the explosive growth of the web data, algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired [8].

Documents pre-processing or dimensionality reduction (DR) allows an efficient data manipulation and representation. Lot of discussions on the pre-processing and DR are there in the current literature and many models and techniques have been proposed. DR is a very important step in text classification, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy and also its tendency to reduce overfitting.

DR techniques can classified into Feature Extraction (FE) [11] and Feature Selection (FS) approaches, as discussed below.

A. Feature Extraction

The process of pre-processing is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors, tokenization, stop words removal, and stemming [10]. FE is the first step of pre processing which is used to presents the text documents into clear word format. So removing stop words and stemming words is the pre-processing tasks [12]. The documents in text classification are represented by a great amount of features and most of them could be irrelevant or noisy [9]. DR is the exclusion of a large number of keywords, base preferably on a statistical

process, to create a low dimension vector [13]. DR techniques have inward much attention recently because effective dimension reduction make the learning task more efficient and save more storage space [14]. Commonly the steps taken please for the feature extractions (Fig.1) are:

Tokenization: A document is treated as a string, and then partitioned into a list of tokens.

Removing stop words: Stop words such as “the”, “a”, “and”... etc are frequently occurring, so the insignificant words need to be removed.

Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form, e.g. connection to connect, computing to compute etc.

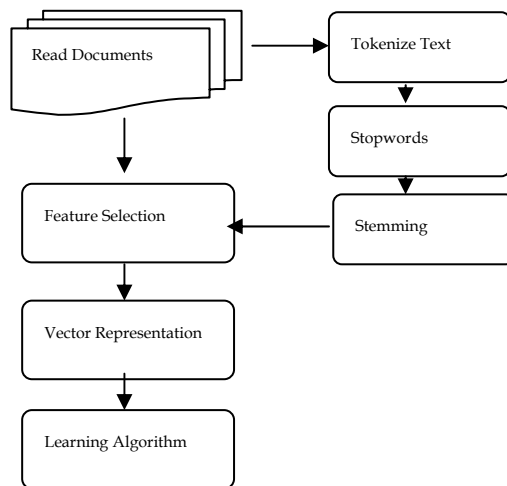


Fig. 1 Document Classification Process

B. Feature Selection

After feature extraction the important step in pre-processing of text classification, is feature selection to construct vector space, which improve the scalability, efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics [15]. The main idea of FS is to select subset of features from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word [9]. The selected features retains original physical meaning and provide a better understanding for the data and learning process [11]. For text classification a major problem is the high dimensionality of the feature space. Almost every text domain has much number of features, most of these features are not relevant and beneficial for text classification task, and even some noise features may sharply reduce the classification accuracy [16]. Hence FS is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

There are mainly two types of feature selection methods in machine learning; wrappers and filters. Wrappers use

the classification accuracy of some learning algorithms as their evaluation function. Since wrappers have to train a classifier for each feature subset to be evaluated, they are usually much more time consuming especially when the number of features is high. So wrappers are generally not suitable for text classification. As opposed to wrappers, filters perform FS independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class [17]. In text classification, a text document may partially match many categories. We need to find the best matching category for the text document. The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weight each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories.

Some of the recent literature shows that works are in progress for the efficient feature selection to optimize the classification process. A novel feature selection method is presented in [17], in which the degrees of deviation from poison distribution are utilized to select informative features. Based on ant colony optimization a new feature selection algorithm is presented in [18], to improve the text categorization. Also in [19] the authors introduced a new weighting method based on statistical estimation of the importance of a word categorization problem. The [20] proposed a new feature scaling method, called class-dependent-feature-weighting (CDFW) using naive Bayes (NB) classifier.

Many feature evaluation metrics have been explored, notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index. Term frequency and document frequency (TF/DF) (Table-1) etc. A good feature selection metric should consider problem domain and algorithm characteristics.

The authors in [21] focused on the document representation techniques and demonstrate that the choice of document representation has a profound impact on the quality of the classifier. They used the centroid-based text classifier, which is a simple and robust text classification scheme, and compare four different types of document representations: N-grams, Single terms, phrases and RDR which is a logic-based documents representation. The N-gram is a string-based representation with no linguistic processing. The Single term approach is based on words with minimum linguistic processing. The phrase approach is based on linguistically formed phrases and single words. The RDR is based on linguistic processing and representing documents as a set of logical predicates. In [22] the authors present significantly more efficient indexing and classification of large document repositories, e.g. to support information retrieval over all enterprise file servers with frequent file updates.

TABLE 1. FEATURE SELECTION TECHNIQUES

Gain Ration	$GR(t_k, c_i) = \frac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(t, c) \log \frac{P(t, c)}{P(t)P(c)}}{- \sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log P(c)}$
Informational Gain(IG)	$IG(w) = - \sum_{j=1}^K P(c_j) \log P(c_j) + P(w) \sum_{j=1}^K P(c_j w) \log P(c_j w) + P(\bar{w}) \sum_{j=1}^K P(c_j \bar{w}) \log P(c_j \bar{w})$ $= H(samples) - H(samples w)$
Chi Square	$\chi^2(f_i, c_j) = \frac{ D \times (\#(c_j, f_i) \#(\bar{c}_j, \bar{f}_i) - \#(c_j, \bar{f}_i) \#(\bar{c}_j, f_i))^2}{(\#(c_j, f_i) + \#(c_j, \bar{f}_i)) \times (\#(\bar{c}_j, f_i) + \#(\bar{c}_j, \bar{f}_i)) \times ((c_j, f_i) + \#(\bar{c}_j, f_i)) \times (\#(c_j, \bar{f}_i) + \#(\bar{c}_j, \bar{f}_i))}$
Conditional mutual Information	$CMI(C S) = H(C) - H(C S_1, S_2, \dots, S_n)$
Document Frequency(DF)	$DF(t_k) = P(t_k)$
Term Frequency(TF)	$tf(f_i, d_j) = \frac{freq_{ij}}{\max_k freq_{kj}}$
Inverse Document Frequency(IDF)	$ idf = \log \frac{ D }{ \#(f_i) }$
Term	$s(t) = P(t \in y t \in x)$
Weighted Ration	$WOddsRation(w) = P(w) \times OddsRatio(w)$
Odd Ration	$OddsRatio(f_i, c_j) = \log \frac{P(f_i c_j)(1 - P(f_i \neg c_j))}{(1 - P(f_i c_j))(P(f_i \neg c_j))}$

C. Semantic and Ontology Base Documents Representation

This section focused on the semantic, ontology techniques, language and the associated issues for documents classification. According to [44] the statistical techniques are not sufficient for the text mining. Better classification will be performed when consider the semantic under consideration. Ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts. It is used to reason about the objects within that domain. Ontology is the explicit and abstract model representation of already defined finite sets of terms and concepts, involved in knowledge management, knowledge engineering, and intelligent information integration [23]. The characteristics of objects and entities (individuals, instances) is a real thing and association (relations) with attribute is used for the titles of the two concepts or entities. Ontology is divided into three categories i.e., Natural Language Ontology (NLO), Domain Ontology (DO) and Ontology Instance (OI) [24]. NLO is the relationship between general lexical tokens of statements based on natural language, DO is the knowledge of a particular domain and OI is the automatically generated web page behaves like an object. Web Ontology Language (OWL) is the ontology support language derived from America DAPRA Agent Markup Language (DAML) and based on ontology, inference and European Ontology Interchange Language (OIL)[25]. OWL claims to be an extension in Resource Description Framework (RDF)[26]. In expressing logical statements because it not only describe classes and properties but also provides

the concepts of namespace, import, cardinality relationship between the classes and enumerated classes. Ontology has been proposed for handling semantically heterogeneity when extracting information from various text sources such as internet [27].

Machine learning algorithms automatically builds a classifier by learning the characteristics of the categories from a set of classified documents, and then uses the classifier to classify documents into predefined categories. However, these machine learning methods have some drawbacks: (1) In order to train classifier, human must collect large number of training text terms, the process is very laborious. If the predefined categories changed, these methods must collect a new set of training text terms. (2) Most of these traditional methods haven't considered the semantic relations between words, so it is difficult to improve the accuracy of these classification methods [6]. (3) The issue of translatability, between one natural language into another natural language. These types of issues identify that machine understanding systems are facing problems. Such issues are discussed in the literature, some of these may be addressed if we have machine readable ontology [32], and that's why this is an important potential area for research.

During the text mining process, ontology can be used to provide expert, background knowledge about a domain. Some recent research shows the importance of the domain ontology in the text classification process, the [27] presents automatic classification of incoming news using hierarchical news ontology, based on this classification on one hand, and on the users' profiles on the other hand,

the personalization engine of the system is able to provide a personalized paper to each user on to her mobile reading device. A novel ontology-based automatic classification and ranking method is represented in [34] where Web documents are characterized by a set of weighted terms, categories are represented by ontology. In [35] the authors presented an approach towards mining ontology from natural language, in which they considered a domain-specific dictionary for telecommunications documents.

How to include user context and preferences in the form of an ontology in order to classify unstructured documents into useful categories and the use of a context-based free text interpreter (CFTI) [36], which performs syntactical analysis and lexical semantic processing of sentences, to derive a description of the content of the unstructured documents, with relevance to the context of the user. In [38] the authors presented a novel text categorization method based on ontological knowledge that does not require a training set. Also an Automatic Document Classifier System based on Ontology and the Naïve Bayes Classifier is proposed in [39].

Ontology's have shown their usefulness in application areas such as knowledge management, bioinformatics, e-learning, intelligent information integration [40], information brokering [41] and natural-language processing [42]. Now it is the positional and challenging area for text classification.

Semantic analysis is the process of linguistically parsing sentences and paragraphs into key concepts, verbs and proper nouns. Using statistics-backed technology, these words are then compared to taxonomy (categories) and grouped according to relevance [43]. Better classification will be performed when consider the semantic under consideration, so the semantically representation of text and web document is the key challenge for the documents classification and knowledge management. Recently many researchers addressed such types of issues.

The authors in [45] present the ambiguity issues in natural language text and present anew technique for resolving ambiguity problem in extracting concept/entity from the text which can improve the document classification process. Multilingual text representation and classification is on of the main and challenging issue in text classification.

In [37] the idea of workflow composition is presented, and addressed the important issues of semantic description of such as services for particular text mining task. Moreover, there are other two open problems in text mining: polysemy, synonymy. Polysemy refers to the fact that a word can have multiple meanings. Distinguishing between different meanings of a word (called word sense disambiguation) is not easy, often requiring the context in which the word appears. Synonymy means that different words can have the same or similar meaning. Some of the natural language issues that should be consider during the text mining process shown in overview [46] is listed below in Table-2.

TABLE 2. SEMANTIC ISSUES FOR DOCUMENTS CLASSIFICATION

Sentence Splitting	How we Identifying sentence boundaries in a document.
Tokenization	How the documents are tokenized and tokens are recorded or annotated, by word or phrase. This is important because many down stream components need the tokens to be clearly identified for analysis.
Part-of-Speech (pos) Tagging	What about the part of speech characteristics and the data annotation. How such components are assigning a pos tag to token pos information.
Stop word list	How stop word list will be taken, and which words are to consider as stop word in which domain.
Stemming	If we reduce the words to their stems, how it will affect the meaning of the documents.
Noisy Data	Which steps are required for the document to be clear from noisy data.
Word Sense	How we clarify the meaning of the word in the text, ambiguity problem.
Collocations	What about the compound and technical terms.
Syntax	How should make a syntactic or grammar analysis. What about data dependency, anaphoric problems.
Text Representation	Which will be more important for representation of the documents: Phrases, Word or Concept and Noun or adjective? And for this which techniques will be feasible to use.
Domain and data understanding for Ontology	How to define the area, data availability and its relation for ontology construction.

Semantically representation of documents is the challenging area for research in text mining. By proper implantation of this will be improve the classification and the information retrieval process.

III MACHINE LEARNING TECHNIQUES

The documents can be classified by three ways, unsupervised, supervised and semi supervised methods. Many techniques and algorithms are proposed recently for the clustering and classification of electronic documents. This section focused on the supervised classification techniques, new developments and highlighted some of the opportunities and challenges using the existing literature. The automatic classification of documents into predefined categories has observed as an active attention, as the internet usage rate has quickly enlarged. From last few years , the task of automatic text classification have been extensively studied and rapid progress seems in this area, including the machine learning approaches such as Bayesian classifier, Decision Tree, K-nearest neighbor(KNN), Support Vector Machines(SVMs), Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation and Genetic Algorithms etc. Normally supervised learning techniques are used for automatic text classification, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labelled documents. Some of these techniques are described below.

A. Rocchio's Algorithm

Rocchio's Algorithm [75] is a vector space method for document routing or filtering in informational retrieval, build prototype vector for each class using a training set of documents, i.e. the average vector over all training document vectors that belong to class c_i , and calculate similarity between test document and each of prototype vectors, which assign test document to the class with maximum similarity.

$$C_i = \alpha * \text{centroid}_{c_i} - \beta * \text{centroid}_{\bar{c}_i} \quad (1)$$

When given a category, the vector of documents belonging to this category is given a positive weight, and the vectors of remaining documents are given negative weight. The positively and negatively weighted vectors, the prototype vector of this category is obtained.

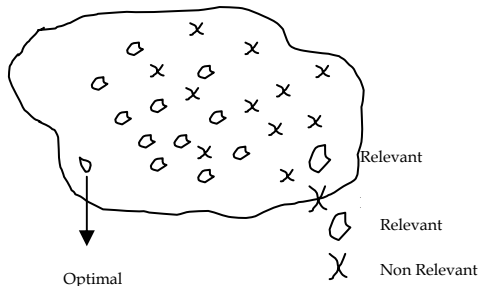


Fig. 2 Rocchio Optimal query for separating relevant and non relevant documents

This algorithm [61] is easy to implement, efficient in computation, fast learner and have relevance feedback mechanism but low classification accuracy. Linear combination is too simple for classification and constant α and β are empirical. This is a widely used relevance feedback algorithm that operates in the vector space model [76]. The researchers have used a variation of Rocchio's algorithm in a machine learning context, i.e., for learning a user profile from unstructured text [77] [78], the goal in these applications is to automatically induce a text classifier that can distinguish between classes of documents.

B. K-nearest neighbor (k-NN)

The k-nearest neighbor algorithm (k-NN) [66] is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. This method is an instant-based learning algorithm that categorized objects based on closest feature space in the training set [62]. The training sets are mapped into multi-dimensional feature space. The feature space is partitioned into regions based on the category of the training set. A point in the feature space is assigned to a particular category if it is the most frequent category among the k nearest training data. Usually Euclidean Distance is typically used in computing the distance between the vectors. The key element of this method is the availability of a similarity measure for identifying neighbors of a particular document [62]. The training phase consists only of

storing the feature vectors and categories of the training set. In the classification phase, distances from the new vector, representing an input document, to all stored vectors are computed and k closest samples are selected. The annotated category of a document is predicted based on the nearest point which has been assigned to a particular category.

$$\arg \max_i \sum_{j=1}^k \text{sim}(D_j | D) * \delta(C(D_j), i) \quad (2)$$

Calculate similarity between test document and each neighbour, and assign test document to the class which contains most of the neighbors. Fig.3.

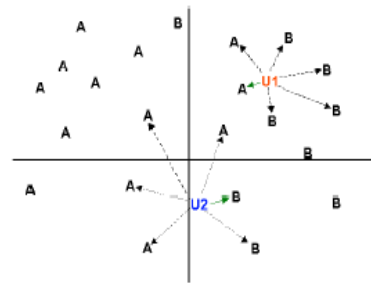


Fig. 3 k-Nearest Neighbor

This method is effective, non parametric and easy to implement. As compare to Rocchio algorithm more local characteristics of documents are considered, however the classification time is long and difficult to find optimal value of k. i.e., to analyze the k-NN and the Rocchio algorithm, some shortcomings of each are identified in [56]. A new algorithm is proposed in [67] which incorporating the relationship of concept-based thesauri into document categorization using a k-NN classifier, while [60] presents the use of phrases as basic features in the email classification problem and performed extensive empirical evaluation using large email collections and tested with three text classification algorithms, namely, a naive Bayes classifier and two k-NN classifiers using TF- IDF weighting and resemblance respectively. The k-nearest neighbor classification method is outstanding with its simplicity and is widely used techniques for text classification. This method performs well even in handling the classification tasks with multi-categorized documents. The major drawback of this method is it uses all features in distance computation, and causes the method computationally intensive, especially when the size of training set grows. Besides, the accuracy of k-nearest neighbor classification is severely degraded by the presence of noisy or irrelevant features.

C. Decision Tree

The decision tree rebuilds the manual categorization of training documents by constructing well-defined true/false-queries in the form of a tree structure. In a decision tree structure, leaves represent the corresponding category of documents and branches represent conjunctions of features that lead to those categories. The well-

organized decision tree can easily classify a document by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf, which represents the goal for the classification of the document.

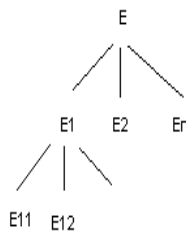


Fig. 4 Decision Tree

The decision tree classification method is outstanding from other decision support tools with several advantages. The main advantage of decision tree is its simplicity in understanding and interpreting, even for non-expert users. Besides, the explanation of a given result can be easily replicated by using simple mathematics algorithms, and provide a consolidated view of the classification logic, which is a useful information of classification.

It can be shown experimentally that text classification tasks frequently involve a large number of relevant features [79]. Therefore, a decision tree's tendency to base classifications on as few tests as possible can lead to poor performance on text classification. However, when there are a small number of structured attributes, the performance, simplicity and understandability of decision trees for content-based models are all advantages. The [80] describe an application of decision trees for personalizing advertisements on web pages.

The major risk of implementing a decision tree is it over fits the training data with the occurrence of an alternative tree that categorizes the training data worse but would categorize the documents to be categorized better [63]. This is due to the classification algorithm of decision tree is made to categorize training data effectively, however neglect the performance of classifying other documents. Besides, huge and excessively complex structure of tree is built from a dataset with very large number of entries.

D. Decision Rules Classification

Decision rules classification method uses the rule-based inference to classify documents to their annotated categories [64] [65]. The algorithms construct a rule set that describe the profile for each category. Rules are typically constructed in the format of "IF condition THEN conclusion", where the condition portion is filled by features of the category, and the conclusion portion is represented with the category's name or another rule to be tested. The rule set for a particular category is then constructed by combining every separate rule from the same category with logical operator, typically use "and" and "or". During the classification tasks, not necessarily every rule in the rule set needs to be satisfied. In the case of handling a dataset with large number of features for each category, heuristics implementation is recommended to reduce the

size of rules set without affecting the performance of the classification. The [49] presents a hybrid method of rule-based processing and back-propagation neural networks for spam filtering. Instead of using keywords, this study utilize the spamming behaviours as features for describing emails.

The main advantage of the implementation of decision rules method for classification tasks is the construction of local dictionary for each individual category during the feature extraction phase [64]. Local dictionaries are able to distinguish the meaning of a particular word for different categories. However, the drawback of the decision rule method is the impossibility to assign a document to a category exclusively due to the rules from different rule sets is applicable to each other. Besides, the learning and updating of decision rule methods need extensive involvement of human experts to construct or update the rule sets. Like the decision trees classification method, the decision rules method does not work well when the number of distinguishing features is large.

E. Naïve Bayes Algorithm

Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' Theorem with strong independence assumptions. A more descriptive term for the underlying probability model would be independent feature model. These independence assumptions of features make the features order is irrelevant and consequently that the present of one feature does not affect other features in classification tasks [99]. These assumptions make the computation of Bayesian classification approach more efficient, but this assumption severely limits its applicability. Depending on the precise nature of the probability model, the naïve Bayes classifiers can be trained very efficiently by requiring a relatively small amount of training data to estimate the parameters necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Due to its apparently over-simplified assumptions, the naïve Bayes classifiers often work much better in many complex real-world situations than one might expect. The naïve Bayes classifiers has been reported to perform surprisingly well for many real world classification applications under some specific conditions [100] [101] [102] [103] [104].

An advantage of the naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters necessary for classification. Bayesian classification approach arrives at the correct classification as long as the correct category is more probable than the others. Category's probabilities do not have to be estimated very well. In other words, the overall classifier is robust enough to ignore serious deficiencies in its underlying naïve probability model.

The main disadvantage of the naïve Bayes classification approach is its relatively low classification performance compare to other discriminative algorithms, such as the

SVM with its outperformed classification effectiveness. Therefore, many active researches have been carried out to clarify the reasons that the naïve Bayes classifier fails in classification tasks and enhance the traditional approaches by implementing some effective and efficient techniques [100] [102] [103] [104] [105].

$$P(c_i | D) = \frac{P(c_i)P(D | c_i)}{P(D)} \quad (4)$$

$$P(D | c_i) = \prod_{j=1}^n P(d_j | c_i) \quad (3)$$

$$\text{Where } P(C_i) = P(C = c_i) = \frac{N_i}{N}$$

$$\text{and } P(d_j | c_i) = \frac{1 + N_{ji}}{M + \sum_{k=1}^M N_{ki}}$$

Naïve Bayes has been one of the popular machine learning methods for many years. Its simplicity makes the framework attractive in various tasks and reasonable performances are obtained in the tasks although this learning is based on an unrealistic independence assumption. For this reason, there also have been many interesting works of investigating naïve Bayes. Recently the [83] shows very good results by selecting Naïve Bayes with SVM for text classification also the authors in [84] prove that Naïve Bayes with SOM give very good results in clustering the documents. The authors in [85] propose a Poisson Naïve Bayes text classification model with weight-enhancing method, and shows that the new model assumes that a document is generated by a multivariate Poisson model. They suggest per-document term frequency normalization to estimate the Poisson parameter, while the traditional multinomial classifier estimates its parameters by considering all the training documents as a unique huge training document. The [86] presented that naïve Bayes can perform surprisingly well in the classification tasks where the probability itself calculated by the naïve Bayes is not important. The authors in a review [87] described that researcher shows great interest in naïve Bayes classifier for spam filtering. So this technique is most widely used in email, web contents, and spam categorization.

Naïve Bayes work well on numeric and textual data, easy to implement and computation comparing with other algorithms, however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences.

F. Artificial Neural Network

Artificial neural networks are constructed from a large number of elements with an input fan order of magnitudes larger than in computational elements of traditional

architectures [106] [107]. These elements, namely artificial neuron are interconnected into group using a mathematical model for information processing based on a connectionist approach to computation. The neural networks make their neuron sensitive to store item. It can be used for distortion tolerant storing of a large number of cases represented by high dimensional vectors.

Different types of neural network approaches have been implemented to document classification tasks. Some of the researches use the single-layer perceptron, which contains only an input layer and an output layer due to its simplicity of implementing [108]. Inputs are fed directly to the outputs via a series of weights. In this way it can be considered the simplest kind of feed-forward network. The multi-layer perceptron which is more sophisticated, which consists of an input layer, one or more hidden layers, and an output layer in its structure, also widely implemented for classification tasks [106].

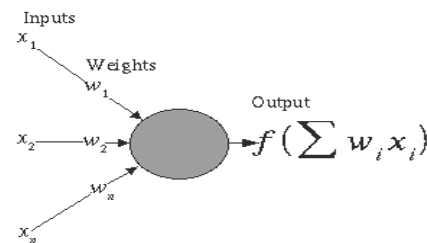


Fig. 5 Artificial Neural Network

The main advantage of the implementation of artificial neural network in classification tasks is the ability in handling documents with high-dimensional features, and documents with noisy and contradictory data. Furthermore, linear speed up in the matching process with respect of the large number of computational elements is provided by a computing architecture which is inherently parallel, where each element can compare its input value against the value of stored cases independently from others [107].

The drawback of the artificial neural networks is their high computing cost which consumes high CPU and physical memory usage. Another disadvantage is that the artificial neural networks are extremely difficult to understand for average users. This may negatively influence the acceptance of these methods.

In recent years, neural network has been applied in document classification systems to improve efficiency. Text categorization models using back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed in [54] for documents classification. An efficient feature selection method is used to reduce the dimensionality as well as improve the performance. New Neural network based document classification method [68], was presented, which is helpful for companies to manage patent documents more effectively.

The ANN can get Inputs x_i arrives through pre-synaptic connections, Synaptic efficacy is modelled using real

weights w_i and the response of the neuron is a nonlinear function f of its weighted inputs.

The output from neuron j for pattern p is O_{pj} where

$$O_{pj}(net_j) = \frac{1}{1 + e^{-\lambda net_j}} \quad (5)$$

and

$$net_j = bias * W_{bias} + \sum_k O_{pk} W_{jk} \quad (6)$$

Neural network for document classification produce good results in complex domains and suitable for both discrete and continuous data (especially better for the continuous domain). Testing is very fast however training is relatively slow and learned results are difficult for users to interpret than learned rules (comparing with Decision tree), Empirical Risk Minimization (ERM) makes ANN try to minimize training error, may lead to overfitting.

G. Fuzzy correlation

Fuzzy correlation can deal with fuzzy information or incomplete data, and also convert the property value into fuzzy sets for multiple document classification [69].

In [55] the authors explore the challenges of multi-class text categorization using one-against-one fuzzy support vector machine with Reuter's news as the example data, and shows better results using one-against-one fuzzy support vector machine as a new technique when compare with one-against-one support vector machine. [61] presented the improvement of decision rule and design a new algorithm of f-k-NN (fuzzy k-NN) to improve categorization performance when the class distribution is uneven, and show that the new method is more effective. So the researchers shows great interest recently to use the fuzzy rules and sets to improve the classification accuracy, by incorporating the fuzzy correlation or fuzzy logic with the machine learning algorithm and the feature selection methods to improve the classification process.

H. Genetic Algorithm

Genetic algorithm [81] aims to find optimum characteristic parameters using the mechanisms of genetic evolution and survival of the fittest in natural selection. Genetic algorithms make it possible to remove misleading judgments in the algorithms and improve the accuracy of document classification. This is an adaptive probability global optimization algorithm, which simulated in a natural environment of biological and genetic evolution, and is widely used for their simplicity and strength. Now several researchers used this method for the improvement of the text classification process. In authors in [82] introduced the genetic algorithm to text categorization and used to build and optimize the user template, and also introduced simulated annealing to improve the shortcomings of ge-

netic algorithm. In the experimental analysis, they show that the improved method is feasible and effective for text classification.

I. Support Vector Machine (SVM)

Support vector machines (SVMs) are one of the discriminative classification methods which are commonly recognized to be more accurate. The SVM classification method is based on the Structural Risk Minimization principle from computational learning theory [109]. The idea of this principle is to find a hypothesis to guarantee the lowest true error. Besides, the SVM are well-founded that very open to theoretical understanding and analysis [110].

The SVM need both positive and negative training set which are uncommon for other classification methods. These positive and negative training set are needed for the SVM to seek for the decision surface that best separates the positive from the negative data in the n -dimensional space, so called the hyper plane. The document representatives which are closest to the decision surface are called the support vector. The performance of the SVM classification remains unchanged if documents that do not belong to the support vectors are removed from the set of training data [99].

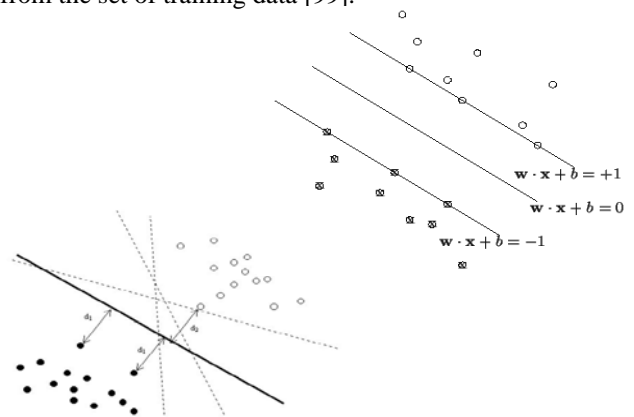


Fig. 6 Illustration of optimal separating hyper plane, hyper planes and support vectors

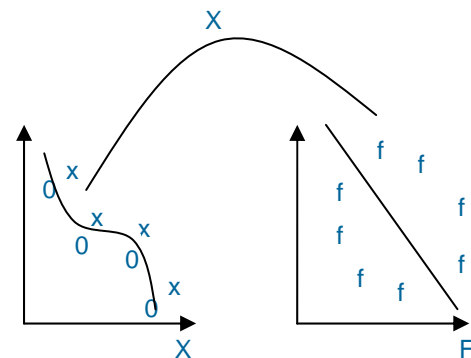


Fig. 7 Mapping non linear input space onto high dimensional space

The SVM classification method is outstanding from the others with its outstanding classification effectiveness [99] [111] [112] [110] [113] [70]. Furthermore, it can handle documents with high-dimensional input space, and culls out most of the irrelevant features. However, the major drawback of the SVM is their relatively complex training and categorizing algorithms and also the high time and memory consumptions during training stage and classifying stage. Besides, confusions occur during the classification tasks due to the documents could be a notated to several categories because of the similarity is typically calculated individually for each category [99].

So SVM is supervised learning method for classification to find out the linear separating hyperplane which maximize the margin, i.e., the optimal separating hyperplane (OSH) and maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplanes are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the neighboring data points of both classes, since in general the larger the margin the lower the generalization error of the classifier.

Maximizing the margin is equivalent to

$$\begin{aligned} \underset{w, b, \zeta_i}{\text{minimize}} \quad & \frac{1}{2} w^T w + C \left(\sum_{i=1}^N \zeta_i \right) \\ \text{subject to} \quad & y_i (w^T x_i - b) + \zeta_i - 1 \geq 0, \quad 1 \leq i \leq N \\ & \zeta_i \geq 0, \quad 1 \leq i \leq N \end{aligned} \quad (9)$$

Introducing Lagrange multipliers α, β , the Lagrangian is:

$$\begin{aligned} \ell(w, b, \zeta_i; \alpha, \beta) &= \frac{1}{2} w^T w + C \sum_{i=1}^N \zeta_i \\ &\quad - \sum_{i=1}^N \alpha_i [y_i (w^T x_i - b) + \zeta_i - 1] - \sum_{i=1}^N \mu_i \zeta_i \\ &= \frac{1}{2} w^T w + \sum_{i=1}^N (C - \alpha_i - \mu_i) \zeta_i \\ &\quad - \left(\sum_{i=1}^N \alpha_i y_i x_i^T \right) w - \left(\sum_{i=1}^N \alpha_i y_i \right) b + \sum_{i=1}^N \alpha_i \end{aligned} \quad (10)$$

The authors in [50] implemented and measured the performance of the leading supervised and unsupervised approaches for multilingual text categorization; they selected support vector machines (SVM) as representative of supervised techniques as well as latent semantic indexing (LSI) and self-organizing maps (SOM) techniques for unsupervised methods for system implementation. In [52] the authors analyses and compares SVM ensembles with four different ensemble constructing techniques, namely bagging, AdaBoost, Arc-X4 and a modified AdaBoost. Twenty real-world data sets from the UCI repository are used as benchmarks to evaluate and compare the performance of these SVM ensemble classifiers by their classification accuracy.

An optimal SVM algorithm via multiple optimal strategies is developed in [47], such as a novel importance weight definition, the feature selection using the entropy weighting scheme, the optimal parameter settings. The SVM is a best technique for the documents classification [83].

IV HYBRID TECHNIQUES

Many new hybrid methods and techniques are proposed recently in the area of Machine Learning and text mining. The concept of combining classifiers is proposed as a new direction for the improvement of the performance of individual classifiers. Recently many methods have been suggested for the creation of ensemble of classifiers. Mechanisms that are used to build ensemble of classifiers [114] include: i) Using different subset of training data with a single learning method, ii) Using different training parameters with a single training method (e.g. using different initial weights for each neural network in an ensemble) and iii) Using different learning methods. [88].

The benefits of local versus global feature sets and local versus global dictionaries in text categorization have examined in [121]. Local features are class dependent features while global features are class independent features. Local dictionaries are class dependent dictionaries while global dictionaries are class independent dictionaries. The best text categorization is obtained using local features and local dictionaries [121].

A New hybrid text document classification approach is proposed in [83], used naive Bayes method at the front end for raw text data vectorization, in conjunction with a SVM classifier at the back end to classify the documents to the right category. They shows that the proposed hybrid approach of the Naive Bayes vectorizer and SVM classifier has improved classification accuracy compared to the pure naive Bayes classification approach. The [84] presents another hybrid method of naïve Bayes with self organizing map (SOM). Proposed Bayes classifier used at the front end, while SOM performs the indexing steps to retrieve the best match cases.

So In the context of combining multiple classifiers for text categorization, a number of researchers have shown that combining different classifiers can improve classification accuracy [89]. It is observed from the Comparison between the best individual classifier and the combined method, that the performance of the combined method is superior [90] [91] [92].

A hybrid method is proposed in [93] in which the learning phase evaluation back propagation neural network (LPEBP) to improve the traditional BPNN. And adopt singular value decomposition (SVD) technique to reduce the dimension and construct the latent semantics between terms, and show that the LPEBP is much faster than the traditional BPNN, which enhances the performance of the traditional BPNN. The SVD technique cannot only greatly reduce the high dimensionality but also enhance the performance. So SVD is to further improve the document classification systems precisely and efficiently.

The [94] a new hybrid technique for text classification is proposed that requires less training data and less computational time and show that text classification that requires fewer documents for training instead of using words, word relation i.e. association rules from these words is used to derive feature set from pre-classified text documents. The concept of Naïve Bayes classifier is then used on derived features and finally only a single concept of genetic presented has been added for final classification.

In[55] the authors explores the challenges of multi-class text categorization using one-against-one fuzzy support vector machine with reuter's news as the example data, and shows better results using one-against-one fuzzy support vector machine as a new technique when compare with one-against-one support vector machine.

A hybrid algorithm is proposed in [56], based on variable precision rough set to combine the strength of both k-NN and Rocchio techniques to improve the text classification accuracy and overcome the weaknesses of Rocchio algorithm.

The authors in [95] suggest a new hybrid approach to web document classification built upon both, graph and vector representations. K-NN algorithm shows that the proposed graph and vector approaches performing better in terms of classification accuracy along with a significant reduction in classification time.

The [96] proposed two methods to modify the standard BPNN and adopt the semantic feature space (SFS) method to reduce the number of dimensions as well as construct latent semantics between terms, and show that the modified methods enhanced the performance of the standard BPNN and were more efficient than the standard BPNN. The SFS method cannot only greatly reduce the dimensionality, but also enhances performance and can therefore be used to further improve text classification systems precisely and efficiently.

The [97] presents a semi-supervised learning method SSRANK for classification task. It leverages the uses of both labelled data and unlabeled data, utilizes views from both traditional IR and supervised learning to conduct data labelling, and relies on a criterion to control the process of data labelling.

A new algorithm of f-k-NN (fuzzy k-NN) proposed in [98] for the improvement of decision rule and design to improve classification performance when the class distribution is uneven, and show that the new method is more effective. The approach of [48] is a nontrivial extension of document classification methodology from a fixed set of classes to a knowledge hierarchy like Gene Ontology.

In [51], the authors proposed a new approach to automatic discovery of implicit rhetorical information from texts based on evolutionary computation methods. In order to guide the search for rhetorical connections from natural-language texts. And in [53], the authors present a segmentation methodology of handwritten documents in their distinct entities, namely, text lines and words.

In [120] the combination of similarity-based learning algorithms and associated thresholding strategies significantly influences the overall performance of text classification. After investigating two similarity-based classifiers (k-NN and Rocchio) and three common thresholding techniques (RCut, PCut, and SCut), they described a new learning algorithm known as the keyword association network (KAN) and a new thresholding strategy (RinS-Cut) to improve performance over existing techniques, and shows that the new approaches give better results.

A new machine learning method is proposed for constructing ranking models in document retrieval [57]. The method, aims to use the advantages of both the traditional Information Retrieval (IR) methods and the supervised learning methods for IR proposed recently.

The main concern of authors in [58] is to investigate the effectiveness of using multi-words for text representation on the performances of text classification. Firstly, a practical method is proposed to implement the multi-word extraction from documents based on the syntactical structure. Secondly, two strategies as general concept representation and subtopic representation are presented to represent the documents using the extracted multi-words. The proposed method launches in [59] for text classification tasks with only unlabeled documents and the title word of each category for learning, and then it automatically learns text classifier by using bootstrapping and feature projection techniques.

V COMPARATIVE STUDY

The growing phenomenon of the textual data needs text mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents. This review focused on the existing literature and explored the documents representation and classification techniques. Text representation is a crucial issue. Most of the literature gives the statistical of syntactic solution for the text representation. However the representation model depend on the informational that we require. Concept base or semantically representations of documents require more attention.

The performance of a classification algorithm in data mining is greatly affected by the quality of data source. Irrelevant and redundant features of data not only increase the cost of mining process, but also degrade the quality of the result in some cases [71]. Each algorithm has its own advantages and disadvantages as described in section II and III.

However, in [6] the author compare the different text classification techniques and have to bear in mind that comparisons are reliable only when based on experiments performed by the same author under carefully controlled conditions. They are instead more problematic when they involve different experiments performed by different authors. In this case various "background conditions," often extraneous to the learning algorithm itself may influence the results. These may include, among others, different

choices in pre-processing (stemming, etc.), indexing, dimensionality reduction and classifier parameter values etc.

A performance comparison in [115] presented a controlled study on a large number of filter feature selection methods for text classification. Over 100 variants of five major feature selection criteria were examined using four well-known classification algorithms: Naive Bayesian (NB) approach, Rocchio-style classifier, k-NN method and SVM system. Two benchmark collections were chosen as the testbeds: Reuters-21578 and small portion of Reuters Corpus Version 1 (RCV1), making the new results comparable to published results. They present that feature selection methods based on χ^2 statistics consistently outperformed those based on other criteria (including information gain) for all four classifiers and both data collections, and that a further increase in performance was obtained by combining uncorrelated and high-performing feature selection methods. The results they obtained using only 3% of the available features are among the best reported, including results obtained with the full feature set. The empirical results of their study suggest that using filter methods which include the χ^2 statistic, combining them with DF or IG, and eliminating the rare words. Such methods were consistently better.

In [116] the authors discussed, that some studies compared feature selection techniques or feature space transformation whereas some others compared the performance of different algorithms. Recently the rising interest towards the Support Vector Machine, various studies showed that SVM outperforms then other classification algorithms. So should we just not problem about other classification algorithms and opt always for SVM? They have decided to investigate this issue and compared SVM to k-NN and naive Bayes on binary classification tasks. An important issue is to compare optimized versions of these algorithms; from their results it shows all the classifiers achieved comparable performance on most problems. One surprising result is that SVM was not a clear winner, despite quite good overall performance. If a suitable pre-processing is used with k-NN, this algorithm continues to achieve very good results and scales up well with the number of documents, which is not the case for SVM. As for Naive Bayes, it also achieved good performance.

The [117] deals with the performance of different classification algorithms and the impact of feature selection algorithm on Logistic Regression Classifier, How it controls False Discovery Rate (FDR) and thus improves the efficiency of Logistic Regression classifier. As per the analysis support vector machine has more parameters than logistics regression and decision tree classifier, SVM has the highest classification precision most of the time, however SVM is very time consuming because of more parameters, demands more computation time. Compared to SVM, logistic regression is computationally efficient. Its results usually have static meaning. However it does not perform well when data set exhibits explicit data structures.

In [118] comparison on four machine learning algorithms, which are Naive Bayesian (NB), neural network (NN), support vector machine (SVM) and relevance vector machine (RVM), are proposed for spam classification. An empirical evaluation for them on the benchmark spam filtering corpora is presented. The experiments are performed based on different training set size and extracted feature size. Experimental results show that NN classifier is unsuitable for using alone as a spam rejection tool. Generally, the performances of SVM and RVM classifiers are obviously superior to NB classifier. Compared with SVM, RVM is shown to provide the similar classification result with less relevance vectors and much faster testing time despite the slower learning procedure, they show that RVM is more suitable than SVM for spam classification in terms of the applications that require low complexity.

In [119] email data was classified using four different classifiers (Neural Network, SVM classifier, Naive Bayesian Classifier, and J48 classifier). The experiment was performed based on different data size and different feature size. The final classification result should be '1' if it is finally spam, otherwise, it should be '0'. This paper shows that simple J48 classifier which make a binary tree, could be efficient for the dataset which could be classified as binary tree.

The [120] shows that two main research areas in statistical text categorization are: similarity-based learning algorithms and associated thresholding strategies. The combination of these techniques significantly influences the overall performance of text categorization. After investigating two similarity-based classifiers (k-NN and Rocchio) and three common thresholding techniques (RCut, PCut, and SCut), they described a new learning algorithm known as the keyword association network (KAN) and a new thresholding strategy (RinSCut) to improve performance over existing techniques. Extensive experiments have been conducted on the Reuters-21578 and 20-Newsgroups data sets, and shows that the new approaches give better results.

Comparing with ANN, SVM capture the inherent characteristics of the data better and embedding the Structural Risk Minimization (SRM) principle which minimizes the upper bound on the generalization error (better than the Empirical Risk Minimization principle) also ability to learn can be independent of the dimensionality of the feature space and global minima vs. local minima. However there are some difficulties in parameter tuning and kernel selection.

VI DISCUSSION AND CONCLUSIONS

This paper provides a review of machine learning approaches and documents representation techniques. An analysis of feature selection methods and classification algorithms were presented. It was verified from the study that information Gain and Chi square statistics are the most commonly used and well performed methods for feature selection, however many other FS methods are

proposed as single or hybrid technique recently, shown good results, and needs more exploration for efficient classification process. Several algorithms or combination of algorithms as hybrid approaches was proposed for the automatic classification of documents, among these algorithms, SVM, NB and kNN classifiers are shown most appropriate in the existing literature.

Most researchers in text classification assume the documents representation as a Bag of Word (BOG), although according to [44] the statistical techniques are not sufficient for the text mining. Text representation is a crucial issue. Most of the literature gives the statistical of syntactic solution for the text representation. However the representation model depend on the informational that we require. Concept base or semantically representation of documents requires more research. Better classification will be performed when consider the semantic under considerations, semantically and ontology base documents representation opportunities were discussed in this paper. With the addition of the ontology and semantic to represent the documents will be more improve accuracy and the classification process. So the identification of features that capture semantic content is one of the important areas for research. The general multiple learning issues in the presence of noise is a tremendously challenging problem that is just now being formulated and will likely require more work in order to successfully develop strategies to find the underlying nature of the manifold.

Several algorithms or combination of algorithms as hybrid approaches were proposed for the automatics classification of documents. Among these algorithms, SVM, NB, kNN and their hybrid system with the combination of different other algorithms and feature selection techniques are shown most appropriate in the existing literature. However the NB is perform well in spam filtering and email categorization, requires a small amount of training data to estimate the parameters necessary for classification. Naive Bayes works well on numeric and textual data, easy to implement comparing with other algorithms, however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences.

SVM classifier has been recognized as one of the most effective text classification method in the comparisons of supervised machine learning algorithms [74]. SVM capture the inherent characteristics of the data better and embedding the Structural Risk Minimization (SRM) principle which minimizes the upper bound on the generalization error (better than the Empirical Risk Minimization principle) also ability to learn can be independent of the dimensionality of the feature space and global minima vs. local minima, however, the SVM has been found some difficulties in parameter tuning and kernel selection.

If a suitable pre-processing is used with k-NN, then this algorithm continues to achieve very good results and scales up well with the number of documents, which is

not the case for SVM [122] [123]. As for naive Bayes, it also achieved good performance with suitable pre-processing. k-NN algorithm performed well as more local characteristic of documents are considered, however the classification time is long and difficult to find optimal value of k.

More works are required for the performance improvement and accuracy of the documents classification process. New methods and solutions are required for useful knowledge from the increasing volume of electronics documents. The following are the some of opportunities of the unstructured data classification and knowledge discovery.

- To improve and explore the feature selection methods for better classification process.
- To reduce the training and testing time of classifier and improve the classification accuracy, precision and recall.
- For Spam filtering and e-mail categorization the user may have folders like electronic bills, e-mail from family, friends and so on, and may want a classifier to classify each incoming e-mail that's automatically move it to the appropriate folder. It is easier to find messages in sorted folders in a very large inbox.
- Automatic allocation of folders to the downloaded articles, documents from text editors and from grid network.
- The use of semantics and ontology for the documents classification and informational retrieval.
- Mining trend, i.e. marketing, business, and financial trend (stock exchange trend) form e-documents (Online news, stories, views and events).
- Stream text require some new techniques and methods for information management.
- Automatic classification and analysis of sentiment, views and extraction knowledge from it. The sentiments and opinion mining is the new active area of text mining.
- Classification and clustering of semi-structured documents have some challenges and new opportunities.
- An implementation of sense-based text classification procedure is needed for recovering the senses from the words used in a specific context.
- Informational extraction of useful knowledge from e-documents and Web pages, such as products and search results to get meaning full patterns.
- To identify or match semantically similar data from the web (that contain huge amount of data and each website represents similar information differently) is an important problem with many practical applications. So web information, integration and schema matching needs more exploration.

REFERENCES

- [1] A. Dasgupta, "Feature selection methods for text classification.", In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 230 -239, 2007.
- [2] Raghavan, P., S. Amer-Yahia and L. Gravano eds., "Structure in Text: Extraction and Exploitation." In. Proceeding of the 7th international Workshop on the Web and Databases(WebDB), ACM SIGMOD/PODS 2004, ACM Press, Vol 67, 2004.
- [3] Oracle corporation, WWW,oracle.com, 2008.
- [4] Merrill lynch, Nov.,2000. e-Business Analytics: Depth Report. 2000.
- [5] Pegah Falinouss "Stock Trend Prediction using News Article's: a text mining approach" Master thesis -2007.
- [6] Sebastiani, F., "Machine learning in automated text categorization" ACM Computing Surveys (CSUR) 34, pp.1 – 47, 2002.
- [7] Andreas Hotho "A Brief Survey of Text Mining" 2005.
- [8] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang Z., "A Novel Feature Selection Algorithm for text categorization." Elsevier, science Direct Expert system with application -2006, 33(1), pp.1-5, 2006.
- [9] Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J., "Measures of Rule Quality for Feature Selection in Text Categorization", 5th international Symposium on Intelligent data analysis , Germeny-2003, Springer-Verlag 2003, Vol2810, pp.589-598, 2003.
- [10] Wang, Y., and Wang X.J., "A New Approach to feature selection in Text Classification", Proceedings of 4th International Conference on Machine Learning and Cybernetics, IEEE- 2005, Vol.6, pp. 3814-3819, 2005.
- [11] Liu, H. and Motoda, ., "Feature Extraction, construction and selection: A Data Mining Perspective.", Boston, Massachusetts(MA): Kluwer Academic Publishers.
- [12] Lee, L.W., and Chen, S.M., "New Methods for Text CategorizationBased on a New Feature Selection Method a and New Similarity Measure Between Documents", IEA/AEI, France 2006.
- [13] Manomaisupat, P., and Abmad k., "Feature Selection for text Categorization Using Self Orgnizing Map", 2nd International Conference on Neural Network and Brain, 2005,IEEE press Vol 3, pp.1875-1880, 2005.
- [14] Yan, J., Liu, N., Zhang, B., Yan, S., Chen, Z., Cheng, Q., Fan, W., and Ma, W., "OCFS: Optimal Orthogonal centroid Feature selection for Text Categorization." 28 Annual International conference on Reserch and Informational retrieval, ACM SIGIR, Barizal, , pp.122-129, 2005.
- [15] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li "An Optimal Svm-Based Text Classification Algorithm" Fifth International Conference on Machine Learning and Cybernetics, Dalian,pp. 13-16 , 2006.
- [16] Jingnian Chen a,b., Houkuan Huang a, Shengfeng Tian a, Youli Qua Feature selection for text classification with Naïve Bayes" Expert Systems with Applications 36, pp. 5432–5435, 2009.
- [17] Hiroshi Ogura, Hiromi Amano, Masato Kondo "Feature selection with a measure of deviations from Poisson in text categorization" Expert Systems with Applications 36, -pp 6826–6832, 2009.
- [18] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri "Text feature selection using ant colony optimization", Expert Systems with Applications 36 pp.6843–6853, 2009.
- [19] P. Scuy, G.W.Mineanu "Beyond TFIDF weighting for text Categorization in the Vector Space Model", 2003.
- [20] E. Youn, M. K. Jeong , "Class dependent feature scaling method using naive Bayes classifier for text datamining" Pattern Recognition Letters , 2009.
- [21] G. Forman, E. Kirshenbaum, "Extremely Fast Text Feature Extraction for Classification and Indexing", Napa Valley California, USA. CIKM'08, October 26–30, 2008
- [22] Mostafa Keikha, Ahmad Khonsari, Farhad Oroumchian, "Rich document representation and classification: An analysis", Knowledge-Based Systems 22 , pp.67–71, 2009.
- [23] D.Fensel, "Ontologies: Silver Bullet for Knowledge Management and e-Commerce", Springer Verlag, Berlin, 2000.
- [24] B. Omelayenko., "learning og ontologies for the Web: the analysis of existent approaches", in the proceeding of the International Workshop on Web Dynamics, 2001.
- [25] OWL Web Ontology Language, viewed March 2008 <http://www.w3.org/TR/owl-features>.
- [26] Sean B. Palmer, "The Semantic Web, an introduction", 2007.
- [27] Lena Tenenboim, Bracha Shapira, Peretz Shoval "Ontology-Based Classification Of News In An Electronic Newspaper" International Conference "Intelligent Information and Engineering Systems" INFOS 2008, Varna, Bulgaria, June-July 2008.
- [28] Lewis, D.D., "Naive (Bayes) at forty The independence assumption in information retrieval", ECML-98, 10th European Conference on Machine Learning, Chemnitz, DE - 1998.
- [29] A. Rafael Calvo, Jae-Moon Lee, Xiabo Li, 'Managin content with automatic document classification", Journal of Digital Information, 5(2) , Article No.282,2004.
- [30] S. Chakrabarti, S. Roy, M. Soundalgekar, "Fast and accurate text classification via multiple linear discriminant projections", International Journal on Very Large Data Bases 12 (2), pp.170–185, 2003.
- [31] Deerwester, S., Dumais, S., Furnas, G.W., Landauer, T.K., Harshman, R; "Indexing by Latent Semantic Analysis". Journal of the Society for Information Science -1990 41 pp. 391-407, 1990.
- [32] Mu-Hee Song, Soo-Yeon Lim, Dong-Jin Kang, and Sang-Jo Lee, "Automatic Classification of Web pages based on the Concept of Domain Ontology", Proc. of the 12th Asia-Pacific Software Engineering Conference, 2005.
- [33] Guiyi Wei, Jun Yu, Yun Ling, and Jun Liu, "Design and Implementation of an Ontology Algorithm for Web Documents Classification", ICCSA 2006, LNCS 3983, pp. 649-658. 2006.
- [34] Jun Fang, Lei Guo, XiaoDong Wang and Ning Yang "Ontology-Based Automatic Classification and Ranking for Web Documents" Fourth International Conference on Fuzzy Systems and Knowledge Discovery -FSKD -2007.
- [35] Alexander Maedche and Ste_en Staab "Mining Ontologies from Text" LNAI 1937, pp. 189-202, 2000. Springer-Verlag Berlin Heidelberg, 2000.
- [36] Ching Kang Cheng, Xiao Shan Pan, Franz Kurfess "Ontology-based Semantic Classification of Unstructured Documents", 2000.
- [37] M. Sarnovský, M. Parali "Text Mining Workflows Construction with Support of Ontologies" 6th International Symposium on Applied Machine Intelligence and Informatics- SAMI 2008.
- [38] Maciej Janik and Krys Kochut "Training-less Ontology-based Text Categorization", 2007.
- [39] Yi-Hsing Chang, Hsiu-Yi Huang "An Automatic Document Classifier System Based On Naïve Bayes Classifier And Ontology" Seventh International Conference on Machine Learning and Cybernetics, Kunming, July 2008.

- [40] G. Wiederhold and M. Genesereth, "The conceptual basis for mediation services", *IEEE Expert / Intelligent Systems*, 12(5):38-47, 1997.
- [41] S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, H.-P. Schnurr, R. Studer, and Y. Sure. "Semantic community web portals", In *Proceedings of the 9th International World Wide Web Conference*, Amsterdam, The Netherlands, May, 15-19, 2000. Elsevier, 2000.
- [42] S. Staab, C. Braun, I. Bruder, A. D'usterhoff, A. Heuer, M. Klettke, G. Neumann, B. Prager, J. Pretzel, H.-P. Schnurr, R. Studer, H. Uszkoreit, and B. Wrenger. *Getess*, "Searching the web exploiting german texts", In *Proceedings of the 3rd international Workshop on Cooperating Information Agents*. Upsala, Sweden, 1999, LNAI 1652, pp. 113-124. Springer, 1999.
- [43] http://www.nstein.com/en/tme_intro.php - 2008.
- [44] Yah, A.s., Hirschman, L., and Morgan, A.A. "Evaluation of text data mining for databassecuration: lessons learned from the KDD challenge cup." *Bioinformatics* 19-(supp.1), pp.i331-i339, 2003.
- [45] H.M.Al Fawareh, S.Jusoh, W.R.S.Osman, "Ambiguity in Text Mining", *IEEE*-2008.
- [46] A.Stavrianou, P. Andritsos, N. Nicoloyannis "Overview and semantic issues of text mining", *SIGMOD Record*, 2007, Vol.36,N03, 2007.
- [47] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li "An Optimal Svm-Based Text Classification Algorithm" *Fifth International Conference on Machine Learning and Cybernetics*, Dalian, 2006.
- [48] H.Kim, and S.S. Chen, "Associative Naïve Bayes Classifier: Automated Linking Of Gene Ontology To Medline Documents" *Pattern Recognition* doi:10.1016/j.patcog.2009
- [49] Chih-Hung Wu , "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks", *Expert Systems with Applications*, pp. 4321–4330, 2009
- [50] Chung-Hong Lee a., Hsin-Chang Yang , "Construction of supervised and unsupervised learning systems for multilingual text categorization", *Expert Systems with Applications*, pp. 2400–2410, 2009.
- [51] John Atkinson a., Anita Ferreira b, Elvis Aravena , "Discovering implicit intention-level knowledge from natural-language texts", *Knowledge-Based Systems* -2009.
- [52] Shi-jin Wang, Avin Mathew, Yan Chen , Li-feng Xi , Lin Ma, Jay Lee, "Empirical analysis of support vector machine ensemble classifiers", *Expert Systems with Applications*, pp. 6466–6476, 2009.
- [53] G. Louloudis, B. Gatos, I. Pratikakis2, C. Halatsis , "Text Line and Word Segmentation of Handwritten Documents", *Pattern Recognition* doi:10.1016/j.patcog.2008.12.016 ,2009
- [54] Bo Yu, Zong-ben Xu, Cheng-hua Li , "Latent semantic analysis for text categorization using neural network", *Knowledge-Based Systems* 21- pp. 900–904, 2008.
- [55] Tai-Yue Wang and Huei-Min Chiang "One-Against-One Fuzzy Support Vector Machine Classifier: An Approach to Text Categorization", *Expert Systems with Applications*, doi: 10.1016/j.eswa.2009.
- [56] Duoqian Miao , Qiguo Duan, Hongyun Zhang, Na Jiao, "Rough set based hybrid algorithm for text classification", *Expert Systems with Applications* -2009 .
- [57] Ming Li, Hang Li , Zhi-Hua Zhou "Semi-supervised document retrieval" *Information Processing and Management* - 2008 .
- [58] Wen Zhang a, Taketoshi Yoshida a, Xijin Tang "Text classification based on multi-word with support vector machine" , *Knowledge-Based Systems* 21 -pp. 879–886, 2008
- [59] Youngjoong Ko a, Jungyun Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques", *Information Processing and Management* 45 -,pp. 70–83, 2009
- [60] Matthew Changa, Chung Keung Poon_, "Using Phrases as Features in Email Classification", *The Journal of Systems and Software* ,doi: 10.1016/j.jss. 2009.
- [61] William W. Cohen and Yoram Singer, "Context-sensitive learning method for text categorization", *SIGIR' 96*, 19th International Conference on Research and Development in Information Retrieval, pp-307-315, 1996.
- [62] Eui-Hong (Sam) Han, George Karypis, Vipin Kumar; "Text Categorization Using Weighted Adjusted k-Nearest Neighbor Classification", *Department of Computer Science and Engineering. Army HPC Research Centre, University of Minnesota, Minneapolis, USA.* 1999.
- [63] Russell Greiner, Jonathan Schaffer; *AIExploratorium - Decision Trees*, Department of Computing Science, University of Alberta, Edmonton, AB T6G 2H1, Canada. 2001. URL :[http://www.cs.ualberta.ca/~aixplore/ learning/ DecisionTrees](http://www.cs.ualberta.ca/~aixplore/learning/DecisionTrees)
- [64] Chidanand Apte, Fred Damerau, Sholom M. Weiss.; "Towards Language Independent Automated Learning of Text Categorization Models", In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 23-30. 1994.
- [65] Chidanand Apte, Fred Damerau, Sholom M. Weiss; "Automated Learning of Decision Rules for Text Categorization", *ACM Transactions on Information Systems (TOIS)*, Vol. 12 , Issue 3, pp. 233 – 251. 1994.
- [66] Tam, V., Santoso, A., & Setiono, R. , "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization", *Proceedings of the 16th International Conference on Pattern Recognition*, pp.235–238, 2002.
- [67] Bang, S. L., Yang, J. D., & Yang, H. J. , "Hierarchical document categorization with k-NN and concept-based thesauri. *Information Processing and Management*", pp. 397–406, 2006.
- [68] Trappey, A. J. C., Hsu, F.-C., Trappey, C. V., & Lin, C.-I., "Development of a patent document classification and search platform using a back-propagation network", *Expert Systems with Applications*, pp. 755–765, 2006 .
- [69] Que, H. -E. "Applications of fuzzy correlation on multiple document classification. Unpublished master thesis", *Information Engineering department, Tamkang University, Taipei, Taiwan*-2000.
- [70] YiMing Yang, Xin Liu; "A Re-examination of Text Categorization Methods, School of Computer Science", *Carnegie Mellon University*. 1999.
- [71] Wu W, Gao Q, Wang M "An efficient feature selection method for classification data mining" *WSEAS Transactions on Information Science and Applications*, 3: pp 2034-2040. 2006.
- [72] Y. Yang; "An evaluation of statistical approaches to text categorization", *Information Retrieval*, Vol.1, No.1, pp. 69-90, 1999.
- [73] T.H. Ng, W.B. Goh, and K.L. Low, "Feature selection, perception learning and a usability case study for text categorization", *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, pp.67-73, 1997.
- [74] Y. Yang, and X. Liu, "An re-examination of text categorization", *Proceedings of the 22nd Annual International ACM*

- SIGIR Conference on Research and Development in Information Retrieval, Berkeley, pp.42-49, August 1999.
- [75] Rocchio, J; "Relevance Feedback in Information Retrieval", In G. Salton (ed.). *The SMART System*: pp.67-88.
- [76] Ittner, D., Lewis, D., Ahn, D; "Text Categorization of Low Quality Images", In: Symposium on Document Analysis and Information Retrieval, Las Vegas, NV .pp. 301-315, 1995
- [77] Balabanovic, M., Shoham Y.: FAB; "Content-based, Collaborative Recommendation", *Communications of the Association for Computing Machinery* 40(3) pp. 66-72, 1997.
- [78] Pazzani M., Billsus, D; " Learning and Revising User Profiles", *The Identification of Interesting Web Sites. Machine Learning* 27(3) pp. 313-331, 1997.
- [79] Joachims, T; "Text Categorization With Support Vector Machines: Learning with Many Relevant Features", In: *European Conference on Machine Learning*, Chemnitz, Germany 1998, pp.137-142 , 1998.
- [80] Kim, J., Lee, B., Shaw, M., Chang, H., Nelson, W; "Application of Decision-Tree Induction Techniques to Personalized Advertisements on Internet Storefronts", *International Journal of Electronic Commerce* 5(3) pp.45-62, 2001.
- [81] Wang Xiaoping, Li-Ming Cao. *Genetic Algorithm Theory, Application and Software[M]*. Xi'an: Xi'an Jiaotong University Press, 2002.
- [82] ZHU Zhen-fang, LIU Pei-yu, Lu Ran, "Research of text classification technology based on genetic annealing algorithm" *IEEE*, 978-0-7695-3311-7/08, 2008.
- [83] Dino Isa, Lam Hong lee, V. P Kallimani, R. RajKumar, "Text Documents Preprocessing with the Bahes Formula for Classification using the Support vector machine", *IEEE, Traction of Knowledge and Data Engineering*, Vol-20, N0-9 pp-1264-1272, 2008.
- [84] Dino Isa., V. P Kallimani Lam Hong lee, "Using Self Organizing Map for Clustering of Text Documents", , Elsevier , *Expert System with Applications*-2008.
- [85] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, "Some Effective Techniques for Naive Bayes Text Classification", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 18, No. 11, , Pp-1457- 1466, November 2006.
- [86] P. Domingos and M. J. Pazzani, "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss," *Machine Learning*, vol. 29, nos. 2/3, pp. 103-130, 1997.
- [87] Thiago S.Guzella, Walimir M. Caminhas " A Review of machine Learning Approches to Spam Filtering", Elsevier , *Expert System with Applications*-2009.
- [88] M. Ikonomakis, S. Kotsiantis, V. Tampakas, "Text Classification Using Machine Learning Techniques", *Wseas Transactions on Computers*, issue 8, volume 4, pp. 966-974, 2005.
- [89] Bao Y. and Ishii N., "Combining Multiple kNN Classifiers for Text Categorization by Reducts", *LNCS* 2534, , pp. 340- 347, 2002.
- [90] Bi Y., Bell D., Wang H., Guo G., Greer K., "Combining Multiple Classifiers Using Dempster's Rule of Combination for Text Categorization", *MDAI*, 2004, 127-138, 2004.
- [91] Sung-Bae Cho, Jee-Haeng Lee, "Learning Neural Network Ensemble for Practical TextClassification", *Lecture Notes in Computer Science*, Volume 2690, Pages 1032- 1036, 2003.
- [92] Nardiello P., Sebastiani F., Sperduti A., "Discretizing Continuous Attributes in AdaBoost for Text Categorization", *LNCS*, Volume 2633, , pp. 320-334, 2003
- [93] "Cheng Hua Li , Soon Choel Park, "An efficient document classification model using an improved back propagation neural network and singular value decomposition" *Expert Systems with Applications* 36 .pp- 3208-3215, 2009.
- [94] S. M. Kamruzzaman and Farhana Haider; "Hybrid Learning Algorithm For Text Classification", 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh.
- [95] Alex Markov and Mark Last, "A Simple, Structure-Sensitive Approach for Web Document Classification", *Springer, AWIC 2005, LNAI 3528*, pp. 293-298, 2005.
- [96] Cheng Hua Li, Soon Cheol Park , "Combination of modified BPNN algorithms and an efficient feature selection method for text categorization.", *Information Processing and Management* 45, 329-340, 2009.
- [97] Ming Li , Hang Li , Zhi-Hua Zhou , "Semi-supervised document retrieval", *Information Processing and Management* 45, pp, 341-355 -2009.
- [98] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin Youli Qu, and Hongbin Dong "An Adaptive Fuzzy kNN Text Classifier", *Springer, ICCS 2006, Part III, LNCS 3993*, pp. 216 - 223, 2006.
- [99] Heide Brücher, Gerhard Knolmayer, Marc-André Mittermayer; "Document Classification Methods for Organizing Explicit Knowledge", *Research Group Information Engineering, Institute of Information Systems, University of Bern, Engehaldenstrasse 8, CH - 3012 Bern, Switzerland*. 2002.
- [100] Andrew McCallum, Kamal Nigam; "A Comparison of Event Models for Naïve Bayes Text Classification", *Journal of Machine Learning Research* 3, pp. 1265-1287. 2003.
- [101] Irina Rish; "An Empirical Study of the Naïve Bayes Classifier", In *Proceedings of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence*. 2001.
- [102] Irina Rish, Joseph Hellerstein, Jayram Thathachar; "An Analysis of Data Characteristics that affect Naïve Bayes Performance", *IBM T.J. Watson Research Center* 30 Saw Mill River Road, Hawthorne, NY 10532, USA. 2001.
- [103] Pedro Domingos, Michael Pazzani; "On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, *Machine Learning*", Vol. 29, No. 2-3, pp.103-130. 1997.
- [104] Sang-Bum Kim, Hue-Chang Rim, Dong-Suk Yook, Huei-Seok Lim; "Effective Methods for Improving Naïve Bayes Text Classification", 7th Pacific Rim International Conference on Artificial Intelligence, Vol. 2417. 2002.
- [105] Susana Eyheramendy, Alexander Genkin, Wen-Hua Ju, David D. Lewis, and David Madigan; "Sparse Bayesian Classifiers for Text Categorization", *Department of Statistics, Rutgers University*. 2003.
- [106] Miguel E. Ruiz, Padmini Srinivasan; "Automatic Text Categorization Using Neural Network", In *Proceedings of the 8th ASIS SIG/CR Workshop on Classification Research*, pp. 59-72. 1998.
- [107] Petri Myllymaki, Henry Tirri; "Bayesian Case-Based Reasoning with Neural Network", In *Proceeding of the IEEE International Conference on Neural Network'93*, Vol. 1, pp. 422-427. 1993.
- [108] Hwee-Tou Ng, Wei-Boon Goh, Kok-Leong Low; "Feature Selection, Perceptron Learning, and a Usability Case Study for Text Categorization, In *Proceedings of the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pp. 67-73. 1997.
- [109] Vladimir N. Vapnik, "The Nature of Statistical Learning Theory", *Springer, New York*. 1995.
- [110] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Fea-

tures" ECML-98, 10th European Conference on Machine Learning, pp. 137-142. 1998.

- [111] Saurav Sahay, "Support Vector Machines and Document Classification" URL: <http://www-static.cc.gatech.edu/~ssahay/sauravsahay7001-2.pdf>
- [112] Soumen Chakrabarti, Shourya Roy, Mahesh V. Soudalgekar, "Fast and Accurate Text Classification via Multiple Linear Discriminant Projection", The International Journal on Very Large Data Bases (VLDB), pp. 170-185. 2003.
- [113] Yi Lin, "Support Vector Machines and the Bayes Rule in Classification", Technical Report No.1014, Department of Statistics, University of Wisconsin, Madison. 1999.
- [114] Wikipedia Ensembles of classifiers, http://en.wikipedia.org/wiki/Ensembles_of_classifiers, 2008.
- [115] Monica Rogati, Yiming Yang "High-Performing Feature Selection for Text Classification" Monica Rogati, Monica Rogati, CIKM'02, November 4-9, 2002, McLean, Virginia, USA., 2002.
- [116] Fabrice Colas and Pavel Brazdil, "Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks", "IFIP International Federation for Information Processing", Springer Boston Volume 217, Artificial Intelligence in Theory and Practice, pp. 169-178, 2006.
- [117] Hanuman Thota, Raghava Naidu Miriyala, Siva Prasad Akula, Mrithyunjaya Rao, Chandra Sekhar Vellanki, Allam Appa Rao, Srinubabu Gedela, "Performance Comparative in Classification Algorithms Using Real Datasets", JCSB/Vol.2 February 2009
- [118] Bo Yu a., Zong-ben Xu b, "A comparative study for content-based dynamic spam classification using four machine learning algorithms", 2008 Elsevier, Knowledge-Based Systems 21, pp. 355-362, 2008.
- [119] Youn and Dennis McLeod, "A Comparative Study for Email Classification, Seongwook Los Angeles", CA 90089, USA, 2006.
- [120] Kang Hyuk Lee, Judy Kay, Byeong Ho Kang, and Uwe Rosebrock, "A Comparative Study on Statistical Machine Learning Algorithms and Thresholding Strategies for Automatic Text Categorization", pp. 444-453, 2002. Springer-Verlag Berlin Heidelberg 2002.
- [121] How, B. C. and Kiong, W. T. (2005). An examination of feature selection frameworks in text categorization. In AIRS. 558-564.
- [122] Pingpeng Yuan, Yuqin Chen, Hai Jin, Li Huang "MSVM-kNN: Combining SVM and k-NN for Multi-Class Text Classification" 978-0-7695-3316-2/08, 2008, IEEE-DOI 10.1109/WSCS.2008
- [123] Fabrice Colas and Pavel Brazdil, "Comparison of svm and some older classification algorithms in text classification tasks", Artificial Intelligence in Theory and Practice (2006), pp. 169-178, 2006.



Aurangzeb Khan received BS-Degree in Computer Science from Gomal University DIKhan, Pakistan and Master Degree in Information Technology From University of Peshawar, Pakistan and is currently a PhD student at the Department of Computer and Information Sciences, Universiti

Teknologi PETRONAS, Malaysia. He is an assistant professor at University of Science and Technology Bannu (USTB) Pakistan. (on study leave). His current research interests include data

mining, sentiment analysis and text classification through AI techniques.



mining, sentiment analysis and text classification through AI techniques.

Baharum Baharudin received his Masters Degree from Central Michigan University, USA and his PhD degree from University of Bradford, UK. He is currently a Senior Lecturer at the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS Malaysia. His research interests lies in Image Processing, Data Mining and Knowledge Management.



mining, sentiment analysis and text classification through AI techniques.

Lam Hong Lee received the bachelor's degree in computer science from University Putra, Malaysia, in 2004 and his PhD from the Faculty of Engineering and Computer Science, University of Nottingham, Malaysia Campus. He is an Assistant Professor at the Faculty of Science, Engineering and Technology of Universiti Tunku Abdul Rahman, Perak Campus, located in Kampar, Malaysia. His current research interest lies in improving text categorization using AI techniques.



mining, sentiment analysis and text classification through AI techniques.

Khairullah Khan received BS-Degree in Computer Science from Gomal University DIKhan, Pakistan and Master Degree in Computer Science from University of Science and Technology Bannu, Pakistan. and is currently a PhD student at the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia. He is Senior Lecturer at University of Science and Technology Bannu (USTB) Pakistan. His current research interests include data mining, opinion mining and text classification through AI techniques

Multilingual Context Ontology Rule Enhanced Focused Web Crawler

Mukesh Kumar and Renu Vig

{mukesh_rai9@yahoo.com,renuvig@hotmail.com}

University Institute of Engineering and Technology, Panjab University, Chandigarh ,INDIA

Abstract— Rapidly growing size and increasing number of Non-English resources on World-Wide-Web poses unprecedented challenges for general purpose crawlers and Search Engines. It is impossible for any search engine to index the complete Web. Focused crawler cope with the growing size by selectively seeking out pages that are relevant to a predefined set of topics and avoiding irrelevant regions of the Web. Rather than collecting and indexing all accessible Web documents, focused crawler analyses its crawl boundary to find the links likely to be the most relevant for the crawl. This paper presents a focused crawler whose crawl strategy is based upon the scores calculated from context ontologies and adaptive classification rules, and which is capable to deal with intermediate multilinguities situations (the situations in which the query language is same as that of target language but the intermediate path may pass through some pages which are written in mixed, in query and some other language, way). It enhances the quality of pages retrieved, because it may be possible that the English meaning of the other language word sequence may itself or point to some pages which are most relevant to the query, and hence should be included in the results, which, yet, are left untouched by all the existing crawlers.

Index Terms— Focused Crawler, Search Engines, Information Retrieval, Ontology, Adaptive Rules

I. INTRODUCTION

The World Wide Web, having more than 350 million pages, continues to grow rapidly at a million pages per day [7]. About 600 MB of text changes every month. Such growth and flux poses basic limits of scale for today's generic crawlers [6] and search engines. It is not possible for any search engines to index the whole Web and to keep track upon the huge consistency management. The only way out of this problem is Focused Crawling [12]. A focused crawler tries to fetch only relevant region of the Web and avoiding irrelevant ones. Since initiation of World Wide Web most of the access is in English language which is the most dominating and preferred one. In recent times there is rapid growth in popularity of internet in semi-English speaking countries like India. Currently the 52% of the whole Web is English and rest is non-English and semi-English. A large fraction of this increasing semi-English population read and writes in mixed way, page written in English and other language (like Hindi).

In this paper focused crawler architecture is presented. The proposed crawler deals with problem of growing size of the Web by focusing its crawl on two parameters:

relevancy score Figure: 1 show a Web page tree in which the black nodes are calculated from context ontologies, and adaptive classification rule score.

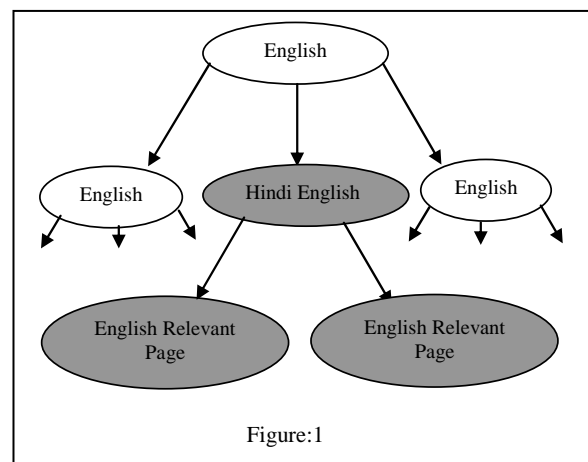


Figure:1

The nodes which are not to be traversed further by any of the existing crawler [2,3,4,5,6,8,9,10,11]. Though the page which is written in Hindi and English, do not contain any relevancy in terms of English text, yet there may be some text written in Hindi whose English transcription makes the page relevant to the user query and that further point to a relevant page written in English. This situation is termed as intermediate multilinguities.

The proposed crawler is able to deal with intermediate multilinguities situations, the situations in which the query language is same as that of target language but the intermediate path may pass through some pages which are written in mixed, in query and some other language, way by making the use of Bilingual dictionary approach.

II. RELATED WORK

The area of multilingual information retrieval has been well explored in the past few decades. The task was approached in two thoughts. One was a translation of the query followed by retrieval in monolingual domain, where as the second was translating the documents into query language and performing retrieval [4]. Broadly, it can be said that the task has been seen as a translation followed by retrieval approach. Bilingual dictionaries derived from Corpus are used for the translation.

Web crawling was simulated by a group of fish migrating on the Web [11]. In the so called fish search, each URL corresponds to a fish whose survivability is dependant on visited page relevance and remote server

No work has yet been done to tackle the intermediate multilinguities situations (Fig.1), which can considerably affect the harvest ratio of the crawler

III. PROPOSED CRAWLER

Fig.2 depicts architecture of the proposed focused Web crawler.It crawler works as per the following code segment:

1. Repeat until Maximum Crawler Limit is reached.
2. Take the user query and initialize the seeds along with their priority as according to the SeedInitializer algorithm discussed later. Set minimum ontology relevance Min_OntRel , and minimum look ahead relevance constant Min_LaRel to some constant values.
3. Download the seed pages as according to their priority and for each seed page go to Step 4.
4. For each link in page go to Step 5.
5. Generate the Context Link with the help of Context Link Extractor and then perform transcription for the intermediate multilinguinity with the help of Multilingual Transcriptor and go to Step 6.
6. Calculate the OBRS and LARS (as discussed later) and put them in the priority queue as according to its OBRS.
7. Retrieve the link with highest OBRS. If its $OBRS > Min_OntRel$ then download the Web page and go to Step 4 Else go to Step 8.
8. If it's $LARS > Min_LaRel$ then download the Web page and go to Step 4.

1. None of them make use of efficient relevance score and tunneling (process of reaching to relevant pages from the irrelevant pages with in the current page) in combination to retrieve the Web documents.
2. Lot of work has been done in general information retrieval, also in multilingual information retrieval, where user enters complete query in one language and results are in some other language, but the crawling is done through the single language only.

Functioning of the main components is discussed below:



A. SEED INITIALIZER

SeedInitializer algorithm work as per the seed detector discussed in [10]. It retrieves the seed URLs from the three most popular search engines Goggle, Yahoo and MSN for the specific keyword. It prioritizes the seeds in the following three classes:

High: URLs occurring across Search Engines more than once.

Medium: URLs repeatedly occurring within the Search Engine, not across the Search Engines.

Low: Other URLs occurring only once within the Search Engine.

B. CRAWLER MANAGER

Crawler manager fetches the URLs from the URL repository and add them to the priority queue. It generates the crawler instances that download the document.

C. CRAWLER

Crawler is a multi-threaded program [10] that is capable of downloading the Web pages from the Web and storing the documents in the document repository. Each crawler has its own queue, which holds the list of URLs to be crawled. The crawler fetches the URLs from the queue. The same or different crawlers would have sent a request to the same server. Busy Server Queue is maintained to have the list of URLs to which the crawlers have sent the request and awaiting for the response. Once the server is connected all the requests are fulfilled. Also instead of disconnecting, it keeps connected for a fixed time interval for the future requests.

D. CONTEXT LINK EXTRACTOR

Context Link Extractor [10] fetches the document from the Document Repository and extracts the URLs. It then checks for the URLs extracted in the URL Fetched. If not found, the surrounding text which include a fixed number of letters preceding and succeeding the hyperlink, the heading or sub heading under which the hyperlink appears is extracted from the document. The extracted link with the context information is passed to Multilingual Transcriptor and URL Repository.

E. MULTILINGUAL TRANSCRIPTOR

This is the component that deals with the intermediate multilinguety situations. It makes use of a bilingual dictionary (e.g. Hindi to English) for transcription. It works as according to the following code segment:

1. If the Context Link is in English language then go to Step 3, Else go to Step 2
2. Locate the Hindi context, remove the noise words and transcript for that in English by making use of the Bilingual Dictionary available, generate the transcribed Context Link, and go to Step 3.
3. Pass the Context Link to the OBRS Generator and LARS Generator.

F. OBRS GENERATOR

Ontology Based Crawling [2, 9] is one of the backbone features of our crawler. OBRS generator makes use of the relevant ontology extracted by the Relevant Ontology Extractor (ROE) and context link passed by the Context Link Extractor (CLE). It uses an Importance Table that importance of each term occurring in the relevant ontology passed from the ROE. A more relevant term to the query will have the more importance and the terms which are common to more than one domain have less importance. Importance Table for a given ontology is given in Table 1.

The following code segment is used for calculation of OBRS for a Context Hyper Link passed from the CLE with the help of Importance Table

Table 1: Importance Table

Ontology terms	Importance
Comp. Sc. And Engg.	1.0
Comp. Engg.	0.9
Comp. Sc.	0.8
Information Tech.	0.5
Computer	0.4
Engineering	0.3

OBRS Generator Algorithm

INPUT: A Context Link (CL) corresponding to a Web page, an Importance Table.

OUTPUT: The relevance score (OBRS) for each Context Link (CL).

Step1: Initialize the relevance score of the Context Link (CL) to 0 i.e. OBRS=0.

Step2: Select first term (T) and corresponding Importance (IMP) from the Importance Table.

Step3: Calculate how many times the term (T) occurs in the Context Link (CL). Let the number of occurrence is calculated in COUNT.

Step4: Multiply the number of occurrence calculated in step 3 with the Importance IMP. Let call this TERM_IMP. And $TERM_IMP = COUNT * IMP$

Step5: Add this term importance to OBRS. So new OBRS will be, $OBRS = OBRS + TERM_IMP$.

Step6: Select the next term and weight from Importance table and go to step3, until all the terms in the weight table are visited.

Step7: End.

G. LARS GENERATOR

Pirkola [1] pointed that for crawling based upon topic it should make use of historical accesses to that particular domain. The proposed crawler makes use of adaptive Rules [3] derived from the classes and link access to improve the crawl and to handle the situation where an irrelevant link in a page may further point to relevant page. For doing this crawler's classifier component is trained with a class and taxonomy, and a set of example documents for each class and call this as train-0 set. Next for each class in the train-0 set we gather all Web pages that the example Web pages in the corresponding class point to. Now we have a collection of class names train-1

set and a set of fetched pages for each class. We know the class distribution of pages to which the documents in each train-0 set class point. For each class in the set train-0, we count the number of referred classes in corresponding train-1 set and generate rules of the form $C_i \rightarrow C_j(P)$, means a page of class C_i can point to a page of class C_j with probability P . P is the ratio of train-1 pages in C_i to all train-1 pages that C_j pages in train-0 refer to. The proposed crawler while seeking Web pages of class C_i attaches priority score P to the pages that it encounters. To demonstrate the approach example is presented. The taxonomy includes four classes:

- Computer Sc. and Engineering (CSE)
- Information Technology (IT)
- Engineering (ENGG)
- Computer (COMP)

Train-0 set can be constructed from any existing directory. Next, for each class, we retrieve the pages that this class's example pages refer to. Assume that we fetch 10 such pages example pages for each class in the train-0 set and that the class distribution among these newly fetched pages i.e train-1 set is as listed in Table 2. Then we can obtain the rules of Table 3.

Adaptive rules may support the tunneling for longer paths using simple application of transitivity among the rules. Also this mechanism is independent of a page's similarity, but rather relies on the probability that a given page's class refer to a target class. This P acts as the Look Ahead Score for the CORE. It can be further manipulated to obtain finer results.

Table 2: Class distribution into train-1 set for each class in train-0 set

<i>CSE</i>	<i>IT</i>	<i>ENGG</i>	<i>COMP</i>
8 URLs for IT	2 URLs for CSE	3 URLs for CSE	10 URLs for COMP
1 URL for ENGG	4 URLs for IT	4 URLs for IT	
1 URL for COMP	4 URLs for ENGG	3 URLs for ENGG	

IV. RESULTS

The proposed crawler is being simulated with 1000 pages and an ontology repository containing ontologies related to all engineering branches under a particular university. And for each test case 0.2 X number of pages out of total X pages are written in mixed way (i.e in English and Hindi), such that the important terms in these 0.2 X pages appears in Hindi. By taking the values of Min_OntRel=5, and Min_LaRel=3 the simulated results are shown in Fig 3. It is a plot of number of relevant pages found against the total number of pages downloaded i.e plot for harvest rate for the baseline focused crawler, non-multilingual context ontology rule enhanced focused web crawler, and multilingual context ontology rule enhanced focused web crawler.

Table 3: Adaptive Rules for the distribution in Table 1, (the number following each rule is the probability P)

<i>CSE</i> : <i>CSE</i> \rightarrow <i>IT</i> (0.8) <i>CSE</i> \rightarrow <i>COMP</i> (0.1) <i>CSE</i> \rightarrow <i>ENGG</i> (0.1)
<i>IT</i> : <i>IT</i> \rightarrow <i>CSE</i> (0.2) <i>IT</i> \rightarrow <i>IT</i> (0.4) <i>IT</i> \rightarrow <i>ENGG</i> (0.4)
<i>ENGG</i> : <i>ENGG</i> \rightarrow <i>CSE</i> (0.3) <i>ENGG</i> \rightarrow <i>IT</i> (0.4) <i>ENGG</i> \rightarrow <i>ENGG</i> (0.3)
<i>COMP</i> : <i>COMP</i> \rightarrow <i>COMP</i> (1.0)

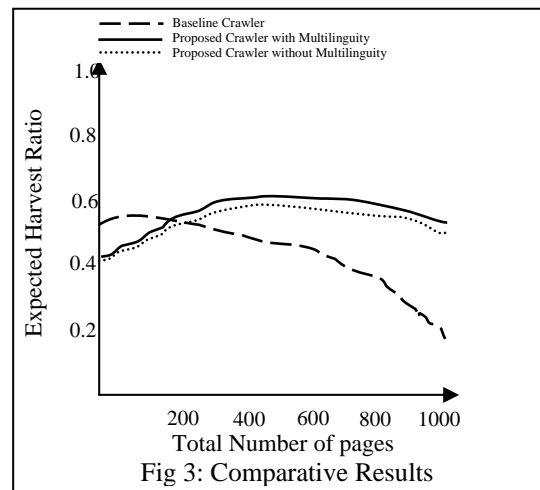


Fig 3: Comparative Results

V. CONCLUSION

A multilingual focused web crawler is presented that makes use of context ontology score and adaptive classification rules for relevancy calculation, and multilingual transcriptor to deal with the intermediate multilinguety. The proposed crawler can deliver improved search results by going through the intermediate multilingual documents. Ontologies served as efficient concepts representation technique. The proposed crawler shows improved harvest ratio when tested for retrieving information for courses run by particular university. It can be used in digital libraries for retrieving documents distributed in the form of documents written in a mixed way. Further work could be done to replace the multilingual transcriptor with multilingual translator for

semantic matching of the intermediate text, which can further enhance the search results.

REFERENCES

- [1]. Ari Pirkola, 2007, "Focused Crawling: A Means To Acquire Biological Data from the Web", VLDB '07, September 23-28, Vienna, Austria, ACM.
- [2]. Debajyoti Mukhopadhyay, Arup Biswas, ukanta Sinha., 2007, "A New Approach to Design Domain Specific Ontology Based Web Crawler", 10th International Conference on Information Technology, IEEE Computer Science, 289-291.
- [3]. Ismail Sengor Altinoglu and Ozgur Ulusoy, November/December 2004, "Exploiting Interclass Rules for Focused Crawling", published in IEEE Intelligent Systems. pp 66-73.
- [4]. Jaime G. Carbonell, Yimming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual information Retrieval: a comparative evaluation. In IJCAI(1), pages 708-715, 1997.
- [5]. J. Cho, H. Garcia-Molina, L. Page. , April 1998 "Efficient Crawling Through URL Ordering" In Proceedings of the 7th International WWW Conference, Brisbane, Australia.
- [6]. Junghoo Cho, Heter Garcia-Molina, WWW 2002 "Parallel Crawlers".
- [7]. K. Bharat and A. Broder. A technique for measuring the relative size and overlap of public web search engines. In Proc. of the 7th WWW Conference 1998.
- [8]. L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Stanford Digital Library Technologies Project.
- [9]. M. Ehrig, A. Maedche., 2003 "Ontology-focused Crawling of Web Documents", In Proceedings of the ACM symposium on Applied computing.
- [10]. M. Yuvrani, N. Ch. S. N. Iyengar, A. Kanan, 2006 "LSCrawler: a Framework for an Enhanced Focused Web Crawler based on Link Structures", in the proceedings of the IEEE/ACM International Conference on web Intelligence.
- [11]. P. M. E. De Bra and R. D. J. Post, "Information retrieval in the World-Wide Web: Making client-based searching feasible", Computer Networks and ISDN Systems. vol. 27, no. 2, pp. 183-192.
- [12]. S. Chakrabarti, M. van den Berg, B. Domc, 1999, "Focused crawling: a new approach to topic-specific Web resource discovery", *Proceedings of the 8th international World Wide Web Conference*, Toronto, Canada.

Integrated Performance and Visualization Enhancements of OLAP Using Growing Self Organizing Neural Networks

Muhammad Usman

Shaheed Zulfikar Ali Bhutto Institute of
Science and Technology, Islamabad,
Pakistan

Email: usmanspak@yahoo.com

Sohail Asghar

Mohammad Ali Jinnah University,
Islamabad, Pakistan

Email: sohail.asghar@jinnah.edu.pk

Simon Fong

University of Macau,
Taipa, Macau SAR

Email: ccfong@umac.mo

Abstract—OLAP performance and its data visualization can be improved using different types of enhancement techniques. Previous research has taken two separate directions in OLAP performance improvement and visualization enhancement respectively. Some recent works have shown the benefits of combining OLAP and Data Mining. Our previous work presents an architecture for the enhancement of OLAP functionality by integrating OLAP and Data Mining. In this paper, we proposed a novel architecture that not only overcomes the existing limitations, but also provides a way for an integrated enhancement of performance and visualization using self organizing neural network. We have developed a prototype and validated the proposed architecture using real-life data sets. Experimental results show that cube construction time and its interactive data visualization capability can be improved remarkably. By integrating enhanced OLAP with data mining system a higher degree of enhancement is achieved which makes significant advancement in the modern OLAP systems.

Index Terms— Clustering, Data Mining, GSOM, Multi-dimensional Data, OLAP, Performance Enhancement, Visualization Techniques

I. INTRODUCTION

OLAP technology refers to a set of data analysis techniques to view the data from all of the transactional

systems in an interactive way in order to support the decision-making process. The fast growing complexity and volumes of the data to be analyzed impose new requirements on OLAP systems [1]. An OLAP system's performance and level of data visualization can be enhanced using different tools and techniques. With the coupling of these enhancement techniques, OLAP functionality can be enhanced [2]. However, OLAP performance improvement and visualization enhancement have been taken separately in the past.

Figure 1 depicts the integration of performance improvement and visualization enhancement practices. Another aspect of enhancement is Data Mining, which aims at the extraction of synthesized and previously unknown insights from large databases [3]. It can be viewed as an automated application of algorithms to detect patterns and extract knowledge from data that is not obvious to the user [4]. Some recent work has shown the benefits of combining OLAP and Data Mining. According to [5], automated techniques of Data Mining can make OLAP more useful and easier to apply in the overall scheme of decision support systems. Furthermore, Data mining techniques like Associations [6], Classification [7], Clustering [8] and Trend Analysis [9] can be used together with OLAP to discover knowledge from data [10].

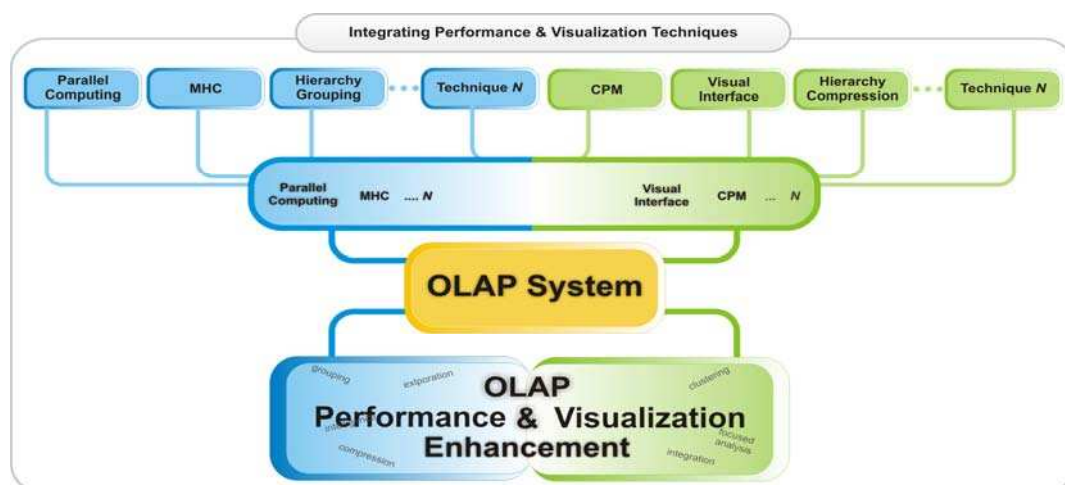


Figure 1. Integration of Enhancement Techniques [2]

In the quest of OLAP enhancement research, Asghar [3] proposed a functionality-enhancement technique using self-organizing neural networks. This technique proposed the integration of Data Mining with OLAP by passing the mined data to the OLAP engine for a more focused analysis and, hence, added intelligence to the OLAP system. The major limitation of the proposed enhancement architecture was the deficiency of work in the enhancement of OLAP performance and visualization. No visualization enhancement technique was used for the expanded view of the OLAP data. Data cube processing time and its physical drive storage were also not discussed. Users of the proposed architecture had to formulate queries to manually retrieve the data of their choice. Interactive visual analysis was missing which is a very attractive functionality of OLAP systems. Typical OLAP data do not change, as they are usually historical data. A major concern is often the support of ad-hoc data exploration by an analyst or other users looking for trends or patterns at v levels of details, perhaps integrated with decision support applications [10].

In this paper, we extend the previous work to overcome the existing limitations and provide an enhanced architecture that can cater for both performance improvement and visualization enhancement. The newly proposed architecture has various modules which allow the integrated performance and visualization enhancement of the OLAP system.

We developed an OLAP prototype system using C# language and other software development tools such as Microsoft SQL server [11], Microsoft Analysis Services [12] and Dundas OLAP services for Windows Form [13]. Experiments are done with data from Forest Cover Type [14] and Zoo [15] data set. It is observed that using the proposed architecture we can enhance the existing OLAP systems in terms of performance and visualization and can get a higher degree of overall enhancement by integrating the performance improvement and visualization enhancement techniques. To the best of our knowledge, no such architecture which features an enhancement solution for OLAP systems for improving performance and enhancing visualization capabilities was ever proposed. Our experimental results show that the cube construction time can be improved remarkably by using the clustered data tables as compared to relational tables. Similarly, by implementing various visualization and enhancement tools and APIs [16] at the OLAP systems can improve the level of interactive data visualization through the use of different types of charts, graphs and data grids.

The remaining of this paper is organized as follows: Section 2 highlights the summary of the past work on which our solution is based. Section 3 addresses the related work. We elaborate the proposed architecture of enhanced OLAP in section 4. The design is followed by description of implementation of our prototype in section 5. Section 6 is where experimental results are discussed and compared with the previous work. A conclusions and possible future research directions are drawn at the end.

II. PREVIOUS WORK

This section provides a brief summary of our previous work about OLAP functionality enhancement, on which our proposed solution is built upon. In our previous work, we extended the capability of the OLAP systems by the use of a neural network. In addition to the usual visualization capabilities, it provided users with the opportunity to analyze data in clusters at different levels of abstraction. The technique used is basically called Growing Self-Organizing Map (GSOM) [17]. GSOM has been developed as a flexible data mining feature mapping method over the traditional Self-Organizing Map (SOM) [18]. The innovation of GSOM is the ability of generating feature maps of different levels of data abstraction using a parameter called 'spread factor'. This spread factor is initially used for generation of hierarchical clusters and analysis technique which is known as dynamic SOM Tree.

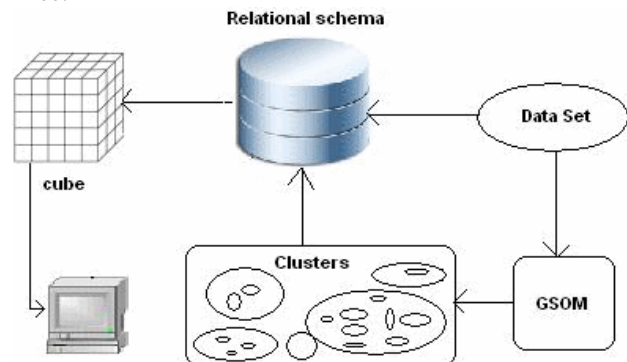


Figure 2. Architecture Proposed in the Previous Work [3].

These hierarchical clusters from the dynamic SOM Tree are subsequently used to provide the OLAP user with the ability to visualize and select data clusters at different levels of abstraction for further detailed analysis. Figure 2 depicts the previously proposed architecture for enhancing OLAP's functionality. This architecture indicates two different approaches to pass the data set to an OLAP engine. The framework we devised and presented in this paper is based on the hierarchical clusters generated by GSOM which are then translated into manual relational tables. The tables are stored in the relational database which serves the data source for the OLAP engine.

The architecture in the past work however has a number of drawbacks. Firstly, the clusters generated from GSOM are to be manually translated into relational tables. This means that user involvement is required for the clusters to be mapped on to a relational schema. Secondly, the OLAP user is unable to perform interactive visualization on the clustered data as there is no such facility available at the front-end. Customized queries are required to view the clustered data which requires knowledge of the complete clustered data in advance. It also lacks of support of Multidimensional expressions (MDX) which is a query language for the OLAP database. Users have to rely on the SQL command, 'GROUP BY' clause to perform runtime aggregation of

data. Cube construction time thus grows unfavorably as the increase of size of the data sets.

The motivation of this paper establishes from the need to keep up with the pace of incremental and fast OLAP development, and the limitations at the previous architecture specifically in terms of its performance and data visualization.

III. RELATED WORKS

As far as performance is concerned in OLAP system, the bottleneck usually involves data processing speeds over the structures of the data cubes. The authors in [5] identified the problem that OLAP operations require complex queries on underlying data, which can be very expensive in terms of computation time. Using parallel computers is one solution. Another approach which is similar to our solution proposed here is to improve cube processing time rather than the OLAP query processing time. For instance, authors in [19] suggested that OLAP performance can be improved by using the (MHC) Multi-dimensional Hierarchical Clustering technique. Clustering was introduced as a way to speed up aggregation queries without additional storage cost for view materialization.

In contrast, our work is to have used GSOM for generating hierarchical clusters for a focused analysis instead of speeding up OLAP queries. Similarly, authors in [20] achieved Heuristic Optimization of OLAP in MHC (multi-dimensionally hierarchically clustered) databases. They found that commercial relational database management systems use multiple one-dimensional indexes to process OLAP queries that restrict multiple dimensions. They presented architecture for MHC databases based on CSB star schema. In our work, we adopted this concept of using relational tables that contain clustered hierarchical information, that are transformed into typical star schema. Along with this concept, researchers in [21] suggested an enhanced OLAP operator based on the Agglomerative Hierarchical Clustering (AHC). The operator is called Operator for Aggregation by Clustering (OpAC) that is able to provide significant aggregates of facts referred to complex objects. Our approach is slightly different that we do not use any operator for clustering. Instead, we use a separate analysis server to construct a cube using the star schema source residing in the database server.

There are other innovated techniques for cube enhancements, such as Cube Presentation Model (CPM) recommended in [22]. CPM can be naturally mapped with an advanced visualization technique called Table Lens. A visual interface for exploring OLAP data with coordinated-dimension hierarchies is introduced in [10].

In literature, a lot of works were devoted to visualization enhancement techniques. Just to name a few, an advanced tool (CommonGIS) for highly interactive visual exploration of spatial data is in [23]. Followed by that, authors in [24] extended a tool for Spatial OLAP, called SOVAT (Spatial OLAP Visualization and Analysis Tool). A hierarchy-driven compression technique for the advanced visualization of

multi-dimensional cubes is suggested in [25]. Then a new visual interactive exploration technique for OLAP is presented in [26]. This is similar to our work in terms of OLAP user facilitation. This enhanced architecture, allows novice users of OLAP technology to explore and analyze OLAP data cubes without sophisticated queries.

Subsequently a framework is proposed in [27] for querying complex multi-dimensional data and transforming irregular hierarchies to make them navigable in a uniform manner. Lately in 2008, Mansmann [28] introduced a comprehensive visual exploration framework which implements OLAP operation as a form of powerful data navigation and allows users to explore data using a variety of interactive visualization techniques. The *Dundas* visualization toolkit which we have used in our work to visualize the OLAP data also allows user to view the data using a number of charts. The use of this software makes our work similar as our choice of visualization of data also provides a number of visualization views to understand and analyze the data in an interactive way.

As observed from our review, research communities advocate that associating Data Mining to OLAP is useful for rich analysis and OLAP tools [21]. By following this fusion idea, we emphasize the coupling of performance and visualization enhancement of OLAP systems as our solution. Why then is there a need for OLAP enhancement architecture? Though a number of enhancement architectures were proposed in the past, none of the work so far was intended towards integrated enhancement of both OLAP performance and visualization, and hence a strong need exists for it [25].

IV. PROPOSED ARCHITECTURE FOR ENHANCED OLAP

To fulfill the growing demands of OLAP users [3], a standardized architecture are required which can easily be deployed as a complete system, it can support the integrated enhancement. Figure 3 depicts such architecture, which enforces integrated enhancement of OLAP's performance and visualization.

We describe the important components of the proposed architecture for the enhancement of OLAP systems. The goal of this architecture is to integrate OLAP with Data Mining and to provide integrated performance and visualization enhancement. To achieve this objective, we deployed separate servers; one is for the database and the other one for the OLAP data. This architecture indicates two channels to pass data to the OLAP engine. The first path is the conventional or non-clustered method where a data set is loaded directly into the database server through Extract, Transform and Load (ETL) process. The data are stored in the form of a relational database. From the relational database a star schema is designed using standard SQL queries and data is loaded into the star schema. The OLAP server takes this star schema as a source to construct OLAP cubes. It also provides storage and management mechanism for the cube data. At the front-end a visualization tool captures the cubes generated by the OLAP server and displays the data in form of charts, reports and tables.

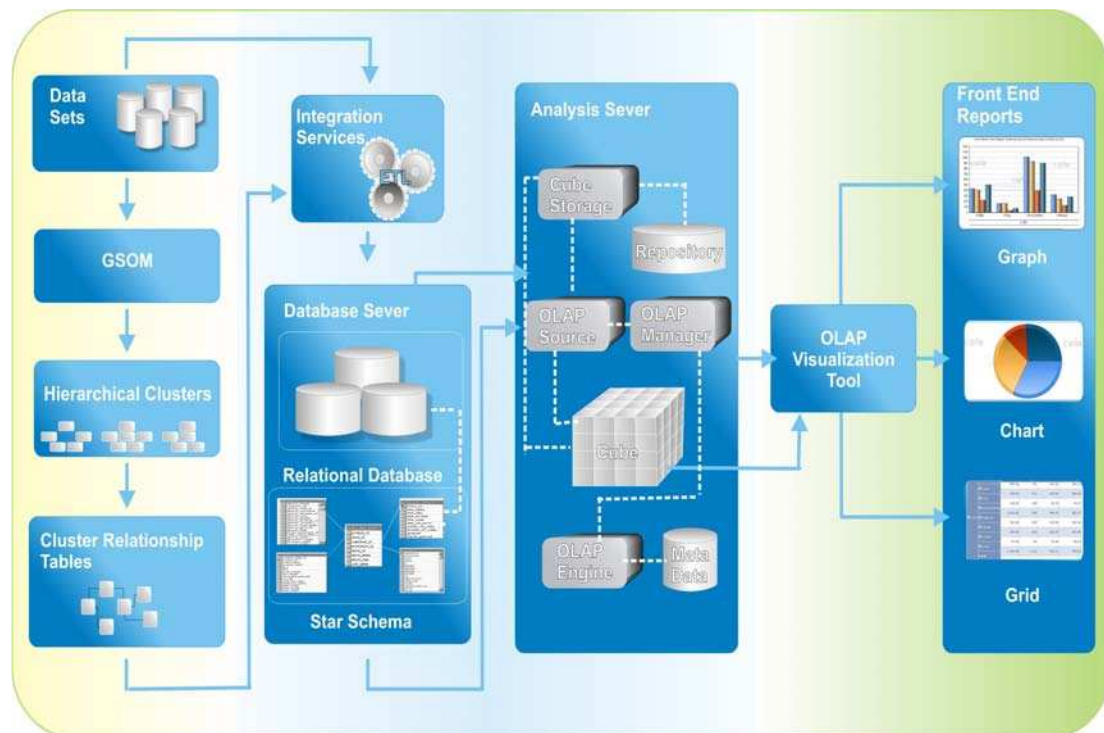


Figure 3. Enhanced OLAP Architecture Proposed.

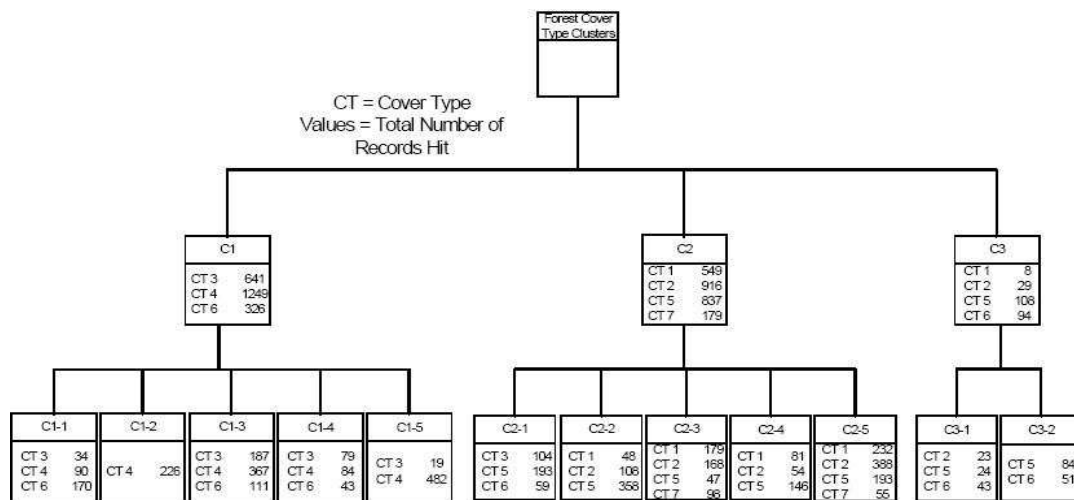


Figure 4. Hierarchical Cluster Decomposition of Forest data set.

A. Data Processing

One unique feature about the architecture we have devised to enhance OLAP is the hierarchical data clusters generated by GSOM. GSOM is based on the design of an unsupervised neural network [29]. The use of GSOM is mainly to produce the hierarchical clusters.

Data set is first fed to GSOM tool which produces the hierarchical clusters using a numerical spread factor. User can set the spread factor to control the number of hierarchical clusters generation using GSOM. An example is given in Figure 4 where the spread factor is 3. Once the clusters are generated the clusters are mapped manually into relational tables. The relational tables are stored in a database in the database server. From these relational tables star schema is created and uploaded with

data. As mentioned previously the star schema becomes the data source. Using this source, cubes of the clustered data are constructed. These clustered cubes become the source of data to be visualized using the prototype in the same manner. From the visualized clusters, user can select the clusters of choice and perform analysis on particular clusters instead of the complete set of data.

B. Visualization

A Front-end visualization tool is connected with the OLAP server that generates cubes using star schema as a data OLAP cubes, displaying the cube data in various views such as charts, graphs, reports and tables at the user's end. Users can perform basic OLAP operations at will using front-end tool; hence, interactive analysis is available using drill down, roll up and other standard

OLAP operations. Visualization capabilities of both traditional and clustered OLAP data are enhanced. The proposed architecture achieves its objective as it integrates OLAP with Data Mining and uses clustered data for performance improvement in terms of cube processing time. In addition to this, the Visualization tool at the front end supports an interactive visual exploration of OLAP data using drill-down, roll-up charts and tables. The visualization tool enhances the visualization capabilities of both traditional and clustered OLAP data.

V. IMPLEMENTATION

This section is presented in two phases: in the first phase we explain the implementation details of the proposed architecture; in the second phase we describe the experiments performed on two different data sets.

Our architecture supports two data loading approaches, clustered and non-clustered. For the non-clustered approach which is the ordinary one, the data set stored in a form of single comma separated text file (depicted in Figure 5) is uploaded in a database as a table using Microsoft SQL Server 2000's Data Transformation Services (DTS) as shown in Figure 6.

Animal	Col001	Col002	Col003	Col004
aardvark	1	0	0	0
antelope	1	0	0	0
bass	0	0	0	1
bear	1	0	0	0
boar	1	0	0	0
buffalo	1	0	0	0
calf	1	0	0	0
carp	0	0	0	1
catfish	0	0	0	1

Figure 5. Data in text file.

After the database has been uploaded with the data, using SQL queries a single *fact table* and three *dimension* tables are created and uploaded with data to form a *star schema* as shown in Figure 7. This star schema residing in the database server is a source data for OLAP server. In this project we are using *Microsoft Analysis Services* as an OLAP data server. The OLAP engine in the Analysis server uses this data source to construct cubes.



Figure 6. Data Transformation Services (DTS) Wizard View.

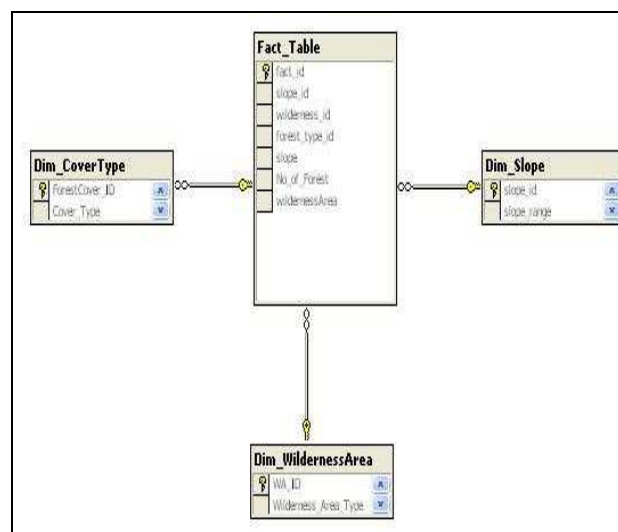


Figure 7. Star Schema of 'Forest Cover Type' Dataset.

Figure 8 shows the cube wizard of MS Analysis Services. Using this wizard user can construct cubes from the data by selecting the different *measures (facts)* available in the *fact table*.

The wizard also facilitates the user to create dimensional hierarchies. Once the cube wizard is finished with the *dimensions* and *measures* selection it prompts the user for the cube storage mode as well. The *Storage Design Wizard* of MS Analysis Services allows for three types of storage modes, namely multidimensional (MOLAP), Relational (ROLAP) and Hybrid (HOLAP). Figure 9 shows the selection interface of storage type.

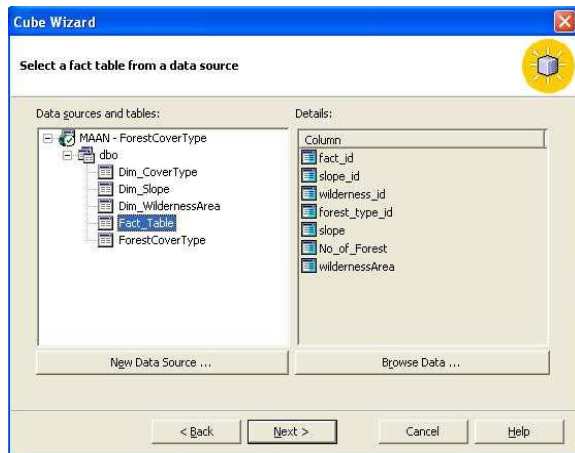


Figure 8. Fact Table Selection Step in the Cube Wizard.

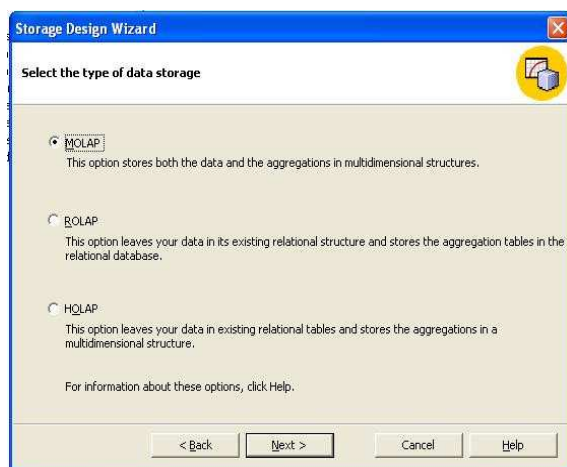


Figure 9. Storage Design Wizard – storage type selection step.

The cube performance can be controlled to some extent by setting the aggregation options of the cube. Aggregations are pre-calculated summaries of data that make querying of the cube faster. The wizard allows users to do setting of the aggregation options as well.

Figure 10 illustrates the aggregate setting option which can be controlled using both performance and size of the cube. Finally, the cube along with all the given settings, are processed and the details are shown to the user.

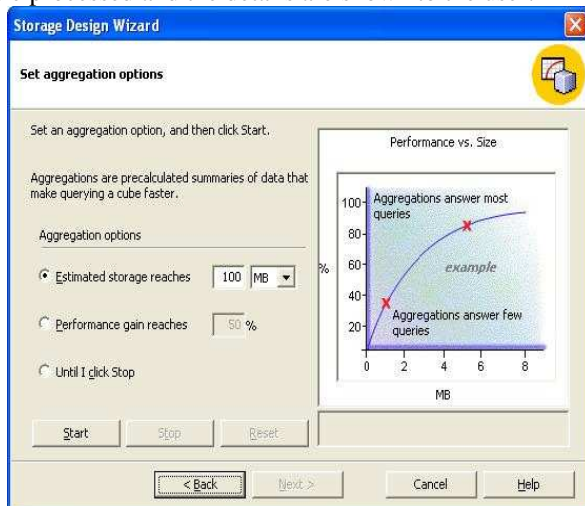


Figure 10. Cube Aggregation Settings –Storage Design Wizard.

Figure 11 shows the summary of the cube process. MS Analysis Services have a built-in *Cube browser* to view the cube data and perform the basic OLAP operations such as *drill-down* and *roll-up*.

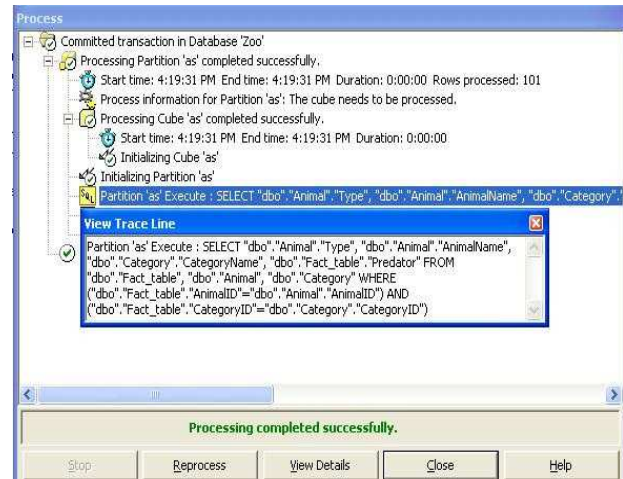


Figure 11. Animal Cube Construction Process Details.

Figure 12 presents a preview of the built-in cube browser of *MS Analysis Services*. The cube browser only shows data in a data grid. In order to enhance the visualization capabilities of the OLAP systems, we have developed a prototype using the *Dundas OLAP services*. With the aid of it, the prototype takes cubes residing in the OLAP server as a source and displays the cube data in the form of a rich and interactive interface. The user can perform the basic OLAP operations and can also see the data in a number of chart types and colorful grids. A number of reports can also be saved in Extensible Markup Language (XML) format.

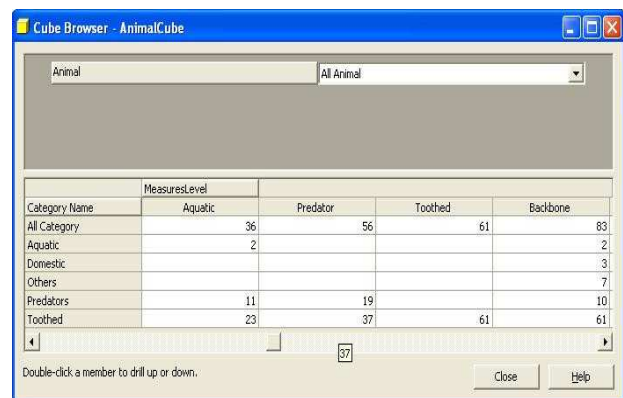


Figure 12. Animal Cube Data in Cube Browser.

Dundas OLAP Services has the following user interface controls: [30].

- OLAP Chart*: A central data visualization area to display the cube data.
- OLAP Grid*: A visualization grid for cube data-analysis.
- OLAP Toolbar*: Quick access to key control functions.
- Cube Selector*: A control to select a cube from a multi-cubed data source.

- e) *Report List*: A collection of reports for storage purposes.
- f) *Dimension Browser*: A control to browse and change the dimensions of the current cube.

Figure 13 shows the user interface of our prototype which has all the above mentioned controls to work on

cube data. The prototype has been developed using *Microsoft visual studio 2005* and *C sharp* programming language.

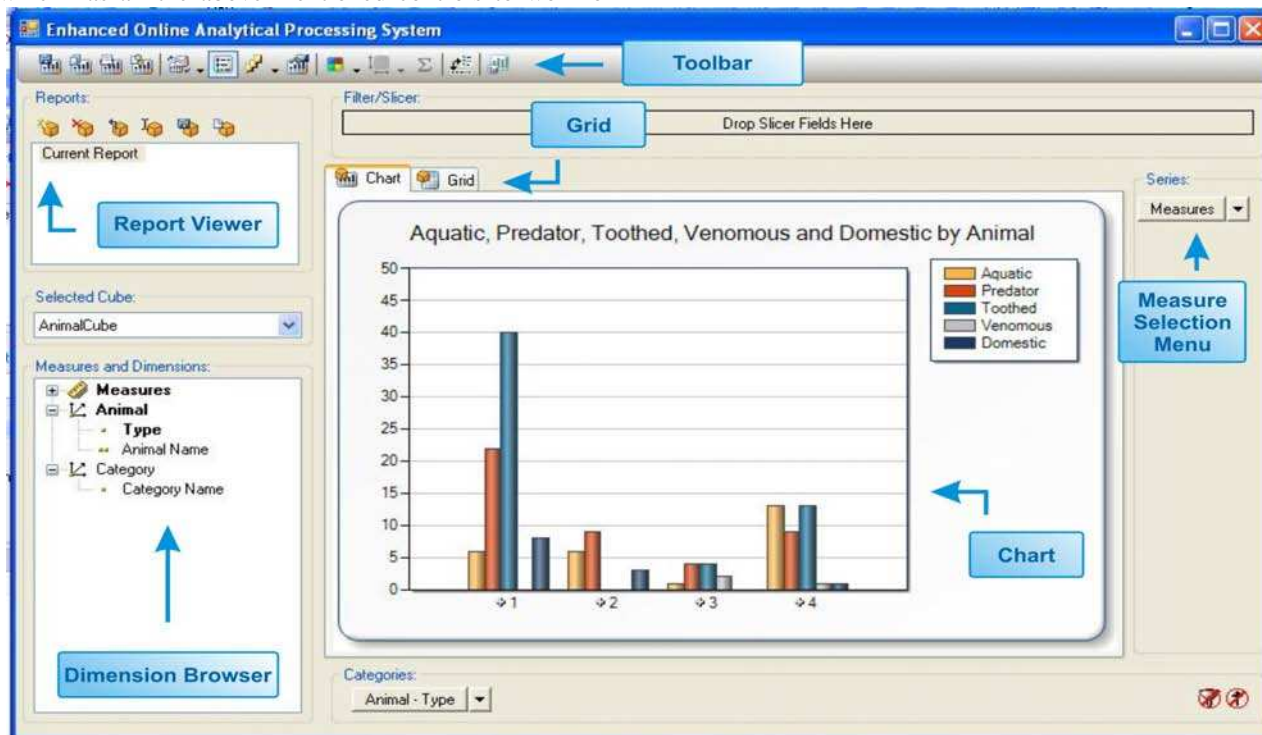


Figure 13. User Interface of Developed Prototype.

The second approach towards OLAP enhancement is the use of GSOM for loading the clustered data into the OLAP system. Data set is first fed into a GSOM tool which produces the hierarchical clusters using the spread factor. User can set the spread factor to control the number of hierarchical clusters generation using GSOM as depicted in Figure 4.

Once the clusters are generated the clusters are mapped manually into relational tables. The relational tables are stored in a database in the database server (MS SQL server 2000). From these relational tables star schema is created and uploaded with data. As mentioned previously the star schema becomes the data source and using this source, cubes of the clustered data are constructed. These cluster-based cubes become the source of data to be visualized using the prototype in the same manner.

Figure 14 shows the cluster-based cube data in the form of bar chart using chart view of the prototype. Using this approach, user can select the clusters of choice and perform analysis on particular clusters instead of the complete cube data.

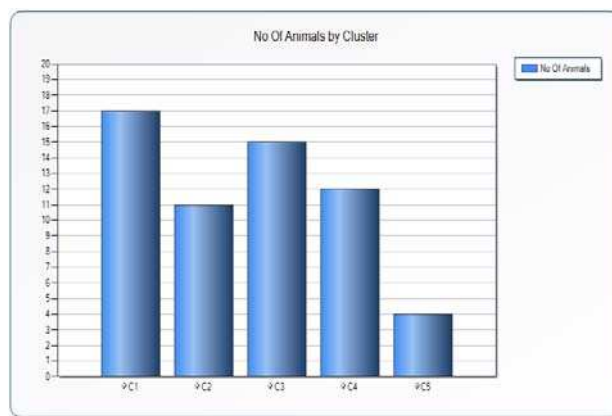


Figure 14. Chart view of clustered zoo data.

VI. EXPERIMENTS

As explained earlier, for the experiments we selected two different data sets for testing and validation of the results of our proposed architecture. The first is a limited dataset called the *Zoo* data set and the second one is a larger data set of *Forest Cover Type*. The idea is to verify how well our new architecture would perform for data of various sizes. Especially we want to demonstrate the integrated enhancements of performance and visualization. We tested both clustered and non-clustered

based approaches on the data sets. However, only the clustered based approach is discussed in details here.

A. Experiment 1 – with Simple Data

In our first experiment we chose a small data set called Zoo Data. Firstly the data set is passed to GSOM to generate the hierarchical clusters. We performed this process using different values of spread factor on the data in order to obtain hierarchical clusters at different levels of abstraction. Only three levels of hierarchical clusters were selected for analysis.

It was observed that the hierarchical clusters generated by using three different values of spread factor such as 0.01, 0.05 and 0.1 identify different groups and subgroups of several animal types. These groups clearly indicate the relevant grouping of data based on user interest and these clusters can be further spread out if necessary. GSOM provides an unbiased grouping of data in terms of clusters. These clusters were picked at diverse levels of abstraction and stored in relational tables. One of the main resulting tables is shown in Table 1.

TABLE I.
CLUSTER HIERARCHY TABLE FOR DATASET ZOO

ID	Clus_Name	Parent Clus	Child Clus	Child-Child Clus
1	Mammals	C1	C1-1	C1-1-1
2	Mammals	C1	C1-2	C1-2-1
3	Mammals	C1	C1-3	C1-3-1
4	Mammals	C1	C1-3	C1-3-2
5	Mammals	C1	C1-4	C1-4-1
6	Mammals	C1	C1-4	C1-4-2
7	Mammals	C1	C1-5	C1-5-1
8	Fish	C2	C2-1	C2-1-1
9	Fish	C2	C2-1	C2-1-2
10	Fish	C2	C2-2	C2-2-1
11	Bird	C3	C3-1	C3-1-1
12	Bird	C3	C3-2	C3-2-1
13	Bird	C3	C3-2	C3-2-2
14	Bird	C3	C3-2	C3-2-3
15	Crawler	C4	C4-1	C4-1-1
16	Crawler	C4	C4-1	C4-1-2
17	Crawler	C4	C4-2	C4-2-1
18	Crawler	C4	C4-3	C4-3-1
19	Insects	C5	C5-1	C5-1-1
20	Insects	C5	C5-1	C5-1-2

From these cluster relationship tables, we created a *fact table*. The two tables output from GSOM, namely Cluster Name Table and Cluster Hierarchy Table are used as *dimensions* to create a *star schema*. The fact table is created and updated using *SQL* queries as shown below.

• Fact Table Creation

```
CREATE TABLE [Fact_table] ( ZooID int
identity(1,1) Not Null, [AnimalID] [int] NULL ,
```

```
[CategoryID] [int] NULL, [Hair] [int] NULL,
[Feathers] [int] NULL, [Airborne] [int]
NULL, [Aquatic] [int] NULL, [Predator]
[int] NULL, [Toothed] [int] NULL, [Backbone]
[int] NULL, [Breathes] [int] NULL, [Venomous]
[int] NULL, [Fins] [int] NULL, [Legs] [int]
NULL, [Tail] [int] NULL,
[Domestic] [int] NULL, [Catsize] [int] NULL
) ON [PRIMARY] GO
```

• Fact Table Data Insertion

```
INSERT INTO [Fact_table] (
[AnimalID],[CategoryID],[Hair],[Feathers],[Eggs],[Mil
k],[Airborne],[Aquatic],[Predator],[Toothed][Backb
one],[Breathes],[Venomous],
[Fins],[Legs],[Tail],[Domestic],[Catsize] )
```

• Fact Table Data Update

```
SELECT
animalid,CategoryID,[Hair],[Feathers],[Eggs],[Mil
k],[Airborne],[Aquatic],[Predator],[Toothed][Backb
one],[Breathes],[Venomous],[Fins],[Legs] ,
[Tail],[Domestic],[Catsize] FROM Animal
a,Category c,Zoo_Data z WHERE
a.AnimalName=z.Animal_Name and
c.CategoryID=z.Category
```

TABLE II.
A FRAGMENT OF FACT TABLE OF CLUSTERED ZOO DATA

fact_id	Clus_id	no_of_animals
1	1	3
2	2	1
3	3	4
...		

The OLAP server takes this *star schema* and uses the clustered data to generate and process a cube. We have named the database having clustered zoo data as *Clustered_Zoo*. We further created an OLAP database and created a new cube called *cluster_zoo_cube*. The OLAP database which has stored the cube becomes the source of cube data. In the prototype, we set the connection so that the front-end tool can connect string with the OLAP server to get cube data and display it in the forms of charts, graphs, grids and reports. The connection string used to connect *Microsoft's Windows Forms* with the OLAP server is as follows;

```
Data Source=[server name]; Provider=msolap.2; Initial
Catalog=Clustered Zoo
```

This connection string identifies the server name on which the OLAP server is hosted. Initial catalog is the OLAP database name which has the cubes stored in it. The prototype uses a function to display the cube data on the front end interface. The function used to display cube data on the client interface is as follow.

```
olapClient1.ShowData();
```

This function is responsible for showing the cube data so that the user can be facilitated with graph, charts and reports. Figure 16 shows the *cluster_zoo_cube* data using the prototype.

By using this interface, the user can drill down and roll up on the cube data and can see an instant manipulation of the chart. Switching from chart view to grid view is easy. Reports can be managed and saved using the report viewer. This interface offers an interactive visualization of charts. The basic operations that can be performed using this prototype are depicted in figures 17, 18 and 19. The bar charts show the hierarchy of the number of animals present in each cluster.

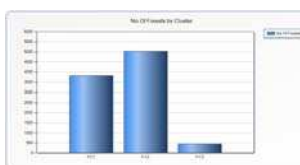


Figure 15. Chart View of Clustered Forest Data Cube.



Figure 16. Drill Down on Cluster 1 (C1)

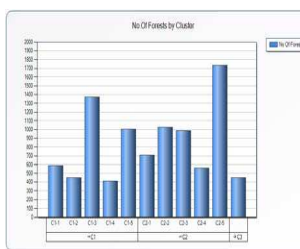


Figure 17. Drill Down on Cluster 2 (C2).

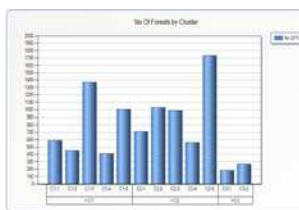


Figure 18. Drill Down on Cluster 3 (C3)

It is obvious from the prototype that the user can see the information of different clusters and can select any cluster of relevant interest for analysis. The presence of this unbiased grouping of data provides the user not only the selection of interested groups for OLAP operations but also to simplify processing of queries.

B. Experiment 2 – with Complex Data

The experiment was conducted over a large data set of *Forest Cover Type* [14]. The data set has 581,012 instances and 54 attributes. The attribute breakdown shows it has 12 measures with 54 columns of data from which 10 are quantitative variables, 4 are binary wilderness areas and 40 binary soil type variables. The data represent a typical analytic situation of certain complexity. The main purpose of the experiment is to validate the capabilities of our proposed architecture. We tested both clustered and non-clustered based approaches on the data sets. The class distribution of the Forest Cover Type data set is shown in table 3.

Firstly, the dataset is fed into the GSOM to generate hierarchical clusters. Hierarchical clusters were then transformed into relational tables. From the relational schema we created a Star Schema. An OLAP database called *Clustered_ForestCoverType* has been created to

store cubes data. The star schema has one fact table and two dimension tables. Using this fact table and dimension tables we created a cube and named it as *Clustered_Forest_Cube*. This cube was stored in the OLAP database and that database was the source for the prototype. The cube was connected to the prototype and the cube data was shown on the front end. Readers are referred to [31] for more details of the process.

TABLE III.
CLASS DISTRIBUTION OF FOREST COVER TYPE [3]

Name	No. of Records
Spruce/Fir	211840
Lodgepole Pine	283301
Ponderosa Pine	35754
Cottonwood Willow	2747
Aspen	9493
Douglas-fir	17367
Krummholz	20510
Total	581012

The Fact table for Forest Cover type data set is in a similar format as in Table 1 in Experiment 1. Using this fact table and dimension tables we created a cube in the same way that was discussed in the previous experiment and named it as *Clustered_Forest_Cube*. This cube was stored in the OLAP database and that database became the source for the prototype. The cube was connected with the prototype and the cube data was shown on the front end. Figure 20 depicts the drill down and roll up operations performed on the clustered cube data showing the number of forests present in each cluster.

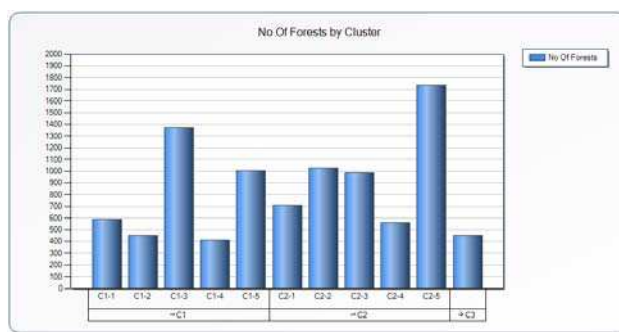


Figure 19. Samples of Drill-Down Operation on Clustered Forest Data.

It is clear from the prototype that the user can perform analysis on the cluster of his/her choice. The visually enriched interface allows interactive exploration of cube data hence enhances the power of OLAP system using the proposed architecture.

VII. DISCUSSION

The results of the experiments are discussed pertaining to the Forest Cover Type data set which represents an example of large size and complex data. This section is divided into two segments. In the first part we discuss the level of performance improvement achieved in terms of cube construction time. The second segment discusses the degree of visualization enhancement for the cube data

A. Performance Improvement

We have performed experiments for both clustered and non-clustered *Forest Cover Type* dataset. The non-clustered data set had originally 581,012 records. This huge amount of data has been loaded in a *MS SQL Server* database called *ForestCoverType*.

From this database a *star schema* has been generated using *SQL* queries, some of which are exposed in Figure 21. This schema became the source for the cube construction. From this non-clustered data set we generated a cube and named it *Forest Cube*, having 3 dimensions and 1 fact table.

```
insert into dbo.Fact_Table(slope_id,wilderness_id,forest_type_id,no_of_forest)
select
slope,Rawah_WA+Neota_WA+Comanche_Peak_WA+Cache_la_Poudre_WA
cover_type,1 from ForestCoverType

update dbo.Fact_Table
set slope_id = 5
where slope_id >60 and slope_id <=75

update dbo.ForestCoverType
set Cache_la_Poudre_WA='4'
where Cache_la_Poudre_WA='1'
```

Figure 20. SQL Queries for star schema generation.

The *Forest Cube* process time was calculated and it was observed that it took 1 second to process 60,180 rows of data to construct the cube.

We carried out the same experiment on the clustered data which was first fed into the GSOM and then transformed into the *Star Schema*. A cube named *Clustered_Forest_Cube* was generated using *Microsoft Analysis Services* depicted in Figure 22.

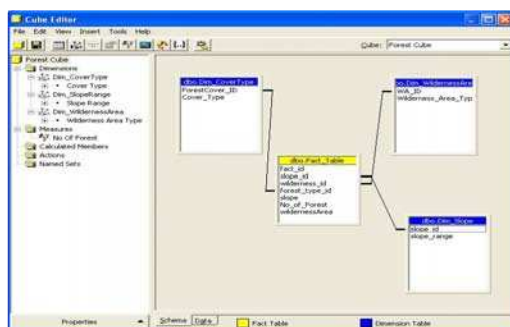


Figure 21. Star schema in Cube Editor of MS Analysis Services.

This clustered data cube was almost instantly processed and its construction time was less than a second. Figure 23 shows the cube construction time comparison of both clustered and non-clustered data.

The results of the experiments to calculate the construction time of a cube shows that the clustered data cube takes less time as compared to the non-clustered data cube. The significant thing about the graph is the rapid variation in the processing time. The processing time line of the non clustered data is increasing rapidly as the volume of data increases. In the case of clustered data the processing time does not increase so sharply. It is identified that if huge amount of data has to be dealt with in the construction of OLAP cube then the clustered

approach is more suitable. For instance, the experiment performed clearly shows that if we increase the size of both clustered and non clustered data it affects the processing time of cube. Interestingly, the rate of increase of processing time when the data is not clustered is quite high. It can be seen in the graph that distance between the two lines keep on increasing as the data increases. For the clustered data the line remains below 200 units of processing time but it reaches up to 1000 units for the same non clustered data.

It is evident from the time comparison graph that the cube processing time can be reduced by using the clustered data, and the user can perform the targeted and fast multidimensional analysis on the clustered cube. Hence, it is shown from our benchmarking experiments that our proposed architecture yields performance improvement of OLAP data cube in computational time.

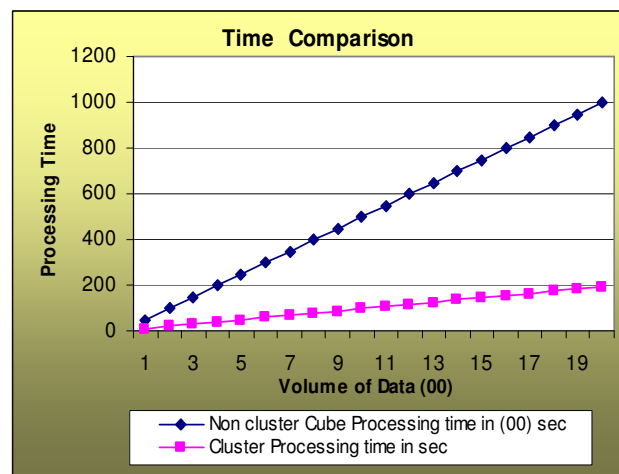


Figure 22. Cube Processing Time Graph.

B. Visualization Improvement

We have constructed cubes for both *Zoo* and *Forest Cover Type* data sets using *Microsoft's Analysis Services* OLAP engine. With the development of the prototype and embedding the OLAP data visualization controls by *Dundas Software*, we provide OLAP users with a rich user interface to perform targeted and interactive visual exploration of the data.

The user can view the same data in a number of charts and graphs simply by selecting the chart type from the *toolbar* in our prototype. For the sake of demonstration, we show the Clustered *Zoo* cube data using our prototype.

Figure 24 shows the outputs of the animal cube in various visual chart formats from our prototype.

Furthermore, users can perform an interactive analysis on the grid as well. Users can drill down and roll up to a level of detail using the "+" and "-" buttons present on the grid and charts. This interactive visualization is enhanced since the previous work only used the *GROUP BY* clause of *SQL queries* to retrieve data from the source system. The comparison of the previous data representation and the representation using the developed prototype is shown in Figure 25.

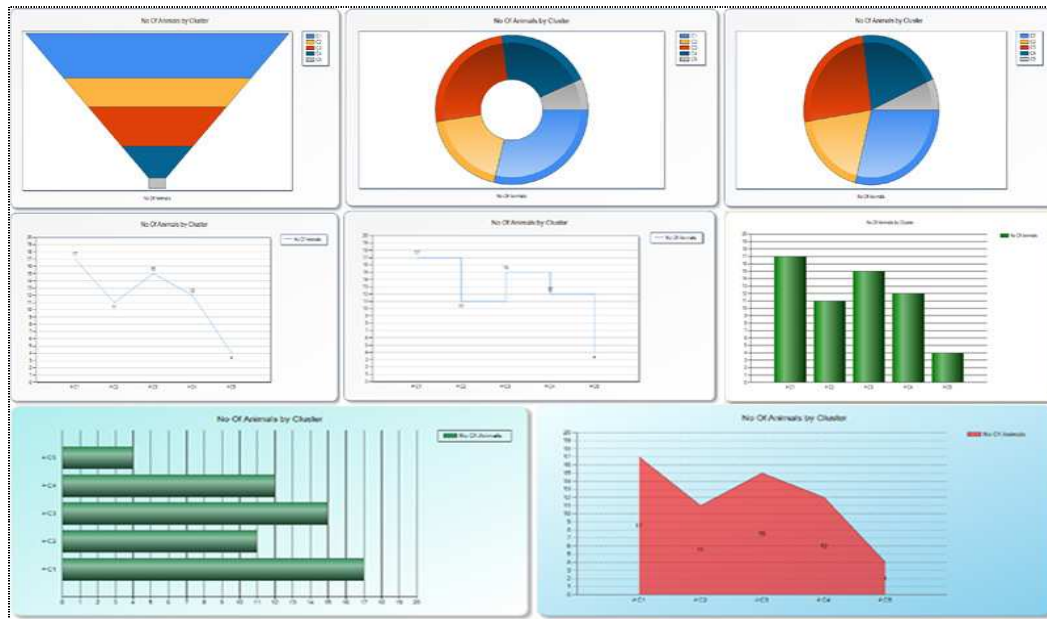


Figure 23. Preview of Different Chart Types Using Our Developed Prototype.

C1	C1-3	C1-3-2	By C1 By C1-3 By C1-3-2	By C1 By C1-3	By C1
<i>Mammals</i>	<i>Domestic Grass Eater</i>	<i>Family Caviidae</i>	2		
				6	
					8

a) Visual representation of cube data in previous work [6]



Chart and Grid View of developed prototype

Figure 24. Comparison of Visual Representation of Data by the Old and New Prototypes.

VIII. CONCLUSION AND FUTURE WORK

In this paper we proposed integrating OLAP performance and visualization enhancement techniques. To support this integration, we devised an architecture for the integrated enhancement using GSOM which generates hierarchical clusters as data input for OLAP. Hierarchical clusters help to enhance targeted analysis in OLAP by views of clusters which is not possible in relational database.

The proposed architecture along with its components is described in details. To demonstrate its advantages, a prototype has been developed and experiments are conducted over two real-life data-sets. It is observed that our architecture is relatively easy to implement and manage. Experimental results show that OLAP performance and visualization can be enhanced significantly. At the end we have compared the cube constructing times with those of the previous work, and

emphasized the benefits of integrating performance improvement with visualization enhancement techniques.

Currently we are working on the dynamic generation of relational tables from the GSOM data. In addition to this we are also working on other Data Mining techniques that can be integrated with OLAP systems to enhance its analysis capabilities.

Furthermore, we are focusing on identifying the other limitations of the current OLAP systems. We are exploring how OLAP can be further extended and enhanced to meet the new challenges and to make it a more effective, efficient and intelligent OLAP system.

REFERENCES

- [1] S. Mansmann and M. Scholl, "Exploring OLAP aggregates with hierarchical visualization techniques," in *Proc. of ACM Symposium on Applied Computing*, 2007, pp. 1067-1073.

- [2] S. Asghar and M. Usman, "Enhancing OLAP performance and visualization," in *Proc. of 6th Int'l Conf. on Information Science Technology and Management (CISTM)*, 2008, pp. 59-70.
- [3] S. Asghar, D. Alahakoon and A. Hsu, "Enhancing OLAP functionality using self-organizing neural networks," *Neural, Parallel and Scientific Computations*, vol. 12, no. 1, pp. 1-20, March 2004
- [4] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *ACM SIGMOD Record*, vol. 26, no. 1, pp. 65-74, March 1997.
- [5] S. Goil and A. Choudhary, "High performance OLAP and data mining on parallel computers," *Data Mining and Knowledge Discovery*, vol. 1, no. 4, pp. 391-417, Dec. 1997.
- [6] J. Hipp, U. Guentzer and G. Nakhaeizadeh, "Algorithms for association ruling mining – a general survey and comparison," *ACM SIGKDD Explorations Newsletter*, vol. 2, no. 1, pp. 58 – 64, June 2000.
- [7] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, no. 1, pp. 63-90, April 1993.
- [8] A. K. Jain, M. N. Murty and P. J. Flynn, "Data clustering: A review," *ACM Computing surveys (CSUR)*, vol. 31, no. 3, pp. 264-323, Sep. 1999.
- [9] M. J. Shaw, C. Subramaniam, G. W. Tan and M. E. Welge, "Knowledge management and data mining for marketing," *Decision Support Systems*, vol. 31, no. 1, pp. 127-137, May 2001.
- [10] M. Sifer, "A visual interface technique for exploring OLAP data with coordinated dimension hierarchies," in *Proc. of 12th ACM Int'l Conf. on Information and Knowledge Management (CIKM)*, 2003, pp. 532-535.
- [11] C. Siedman, *Data Mining with Microsoft SQL Server 2000 Technical Reference*, Microsoft Press, Redmond, WA, 2001.
- [12] S. Soni and W. Kurtz, "Analysis services: Optimizing cube performance using Microsoft SQL server 2000 analysis services," *Microsoft SQL Server 2000 Technical Articles* March 2001.
- [13] "Dundas Chart for ASP.NET," Help Document, [Online], <http://support.dundas.com/OnlineDocumentation/WebChart2005>.
- [14] J. A. Blackard, D. J. Dean and C. W. Anderson, Forest cover type, *The UCI KDD Archive* [<http://kdd.ics.uci.edu>]. Irvine, CA: University of California, Department of Information and Computer Science (1998).
- [15] E. Keogh, C. Blake and C. J. Merz, "UCI repository of machine learning database," 1999. [Online]. [http: / /www.ics.uci.edu /~mllearn /MLRepository.html](http://www.ics.uci.edu/~mllearn/MLRepository.html)
- [16] T. Imielinski, A. Virmani and A. Abdulghani, "DataMine: Application programming interface and query language for database mining," in *Proc. of 2nd KDD Conf.*, 1996, pp. 256-262.
- [17] D. Alahakoon, S. K. Halgamuge and B. Srinivasan, "Dynamic self organising maps with controlled growth for knowledge discovery," *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 601-614, May 2000.
- [18] T. Kohonen, *Self-Organising Maps*, 3rd ed. Springer-Verlag, Berlin, 2001.
- [19] V. Markl, F. Ramasak and R. Bayer, "Improving OLAP performance by multi-dimensional hierarchical clustering," in *Proc. of 1999 Int'l Symposium on Database Engineering and Applications*, 1999, p. 165.
- [20] D. Theodoratos and A. Tsois, "Heuristic optimization of OLAP queries in multidimensionally hierarchically clustered databases," in *Proc. of 4th ACM Int'l Workshop on Data Warehousing and OLAP*, 2001, pp. 48-55.
- [21] R. B. Messaoud, O. Boussaid and S. Rabaseda, "A new OLAP aggregation based on the AHC technique," in *Proc. of 7th ACM Int'l Workshop on Data Warehousing and OLAP (DOLAP)*, 2004, pp. 65-72.
- [22] A. S. Maniatis, P. Vassiliadis, S. Skiadopoulos and Y. Vassiliou, "Advanced visualization for OLAP", in *Proc. of 6th ACM Int'l Workshop on Data Warehousing and OLAP (DOLAP)*, 2003, pp. 9-16.
- [23] A. Voss, V. Hernandez, H. Voss and S. Scheider, "Interactive visual exploration of multidimensional data: Requirements for common GIS with OLAP," in *Proc. of 15th Int'l Workshop on Database and Expert Systems Applications (DEXA)*, 2004, pp. 883-887.
- [24] M. Scotch and B. Paramanto, "SOVAT: Spatial OLAP visualization and analysis tool," in *Proc. of 38th Annual Hawaii Int'l Conf. on Systems Sciences (HICSS)*, 2005, p. 165.
- [25] A. Cuzzocrea, D. Sacca and P. Serafino, "A hierarchy driven compression technique for advanced OLAP visualization of multidimensional data cubes", in *Proc. of 8th Int'l Conf. on Data Warehousing and Knowledge Discovery (DaWak'06)*, Springer Verlag 2006, pp. 106-119.
- [26] K. Techapichetvanich and A. Datta, "Interactive visualization for OALP," in *Int'l Conf. on Computational Science and its Applications (ICCSA)*, 2005, pp. 206-214.
- [27] S. Mansmann and M. Scholl, "Extending visual OLAP for handling irregular dimensional hierarchies," in *Proc. of 8th Int'l Conf. on Data Warehousing and Knowledge Discovery (DaWak'06)*, Springer Verlag 2006, pp. 95-105.
- [28] S. Mansmann and M. Scholl, "Visual OLAP: A new paradigm for exploring multidimensional aggregates," in *Proc. of IADIS Int'l Conf. on Computer Graphics and Visualization (CGV)*, 2008, pp. 59-66.
- [29] B. Fritzke, "Growing cell structures - a self organizing network for unsupervised and supervised learning," *Neural Networks*, vol. 7, no. 9, pp. 1441-1460, 1994.
- [30] Dundas Chart for ASP.NET - OLAP Services, Overview of Dundas OLAP Architecture, 2005 - 2009 Dundas Data Visualization, Source:[http://support.dundas.com/OnlineDocumentation/ WebOLAP/Overview%20of%20Dundas%20OLAP%20Ar chitecture.html](http://support.dundas.com/OnlineDocumentation/WebOLAP/Overview%20of%20Dundas%20OLAP%20Architecture.html). [Online]. [Accessed: Dec. 2008].
- [31] M. Usman, S. Asghar, S. Fong, "An Architecture of Integrated Enhancement of OLAP Using Growing Self Organizing Neural Networks", *International Conference on Computer Engineering and Systems*, Cairo, Egypt, IEEE, submitted.

Design and Implementation of an Online Social Network with Face Recognition

Ray K. C. Lai

University of Macau, Av. Padre Tomás Pereira, Taipa, Macao, China

Email: chonchondotcom@gmail.com

Jack C. K. Tang, Angus K.Y. Wong and Philip I.S. Lei

Macao. Polytechnic Institute, Rua de Luis Gonzaga Gomes, Macao, China

Email: { jacktang, kywong, philiplei }@ipm.edu.mo

Abstract—This paper presents the idea of an online social network that makes use of the face recognition technology. With the technology, friend relationship can be established without the need of having the text-based information of a user, friend recommendation algorithm can be more accurate, and face tagging can be done automatically. The design and implementation issues of such system will be discussed. The prototype of our proposed social network will be demonstrated to show the feasibility of adopting the face recognition technology in online social networks.

Index Terms—face recognition, face tagging, social networks

I. INTRODUCTION

Social network is a set of people (or organizations or other social entities) connected by a set of socially-meaningful relationships [1]. Online social network has become very popular in recent years. Some popular social networking websites have hundreds of million users registered. In this kind of sites, users can update their personal profiles, notify friends about themselves, play game and share photos with their friends internationally.

Before interacting with friends, a user must add them to form a relationship. In the existing social networking web sites, it is found that, to find a friend, the web sites are commonly based on the text-based information such as email addresses, names of friends, school names of friends, etc. Though this approach is working, in this paper, we argue that face recognition can be used to improve the friend searching and other services in online social network.

Research in automatic face recognition can date back

at least until the 1960s [2]. However, most face recognition algorithms were developed in 1980s and 1990s. Two most common of them are Principal Component Analysis (PCA) and Independent Component Analysis (ICA). Kirby and Sirovich were among the first to apply PCA to face images, and showed that PCA is an optimal compression scheme that minimizes the mean squared error between the original images and their reconstructions for any given level of compression [3][4]. Turk and Pentland popularized the use of PCA for face recognition [5]. PCA were matched the images in the database by projecting them onto the basis vectors and finding the nearest compressed image in the subspace (called eigenspace). ICA can also be used to create feature vectors that uniformly distribute data samples in subspace [6][7]. This conceptually very different use of ICA produces feature vectors that are not spatially localized. Instead, it produces feature vectors that draw fine distinctions between similar images in order to spread the samples in subspace.

Though the face recognition technology has been well established, there are no online social networks using it. The technology can provide a number of advantages to online social networks, including:

1. Friends on photos can be searched even without text-based information. It is particular useful when we only took photo with a new friend but forgot to exchange the contact.
2. The friend recommendation algorithm can be more accurate because it can make use both the text-based and face information.
3. Face tagging can be automatically done with the face recognition technology.

In the rest of this paper, we will show the design and architecture of our proposed online social network with face recognition functions. After that, we will reveal how the face recognition technology can be used to design a new friendship search algorithm, to generate friend recommendation list, and to design photo-based search and match functions. Finally, we will discuss the implementation of the system and demonstrate our prototype social network.

Manuscript received October 22, 2009; revised December 29, 2009.

Ray K. C. Lai is a research master student at University of Macau, Av. Padre Tomás Pereira, Taipa, Macao, China (e-mail: chonchondotcom@gmail.com).

Jack C. K. Tang is a research student at Macao Polytechnic Institute, Rua de Luis Gonzaga Gomes, Macao (e-mail:jacktang@ipm.edu.mo).

Angus K.Y. Wong is an associate professor at Macao. Polytechnic Institute, Rua de Luis Gonzaga Gomes, Macao (e-mail: kywong@ipm.edu.mo).

Philip I.S. Lei is an associate professor at Macao. Polytechnic Institute, Rua de Luis Gonzaga Gomes, Macao (e-mail: philiplei@ipm.edu.mo).

TABLE I.
COMPARISON OF ONLINE SOCIAL NETWORKS WITH AND WITHOUT FACE
RECOGNITION TECHNOLOGY

	Existing social networks	Social networks using face recognition technology
Forming a relationship	The personal information of a friend is needed	Any face of a friend appearing in any pictures of the user is sufficient
Searching and recommending a friend	Using text-based search algorithm	Using face-recognition algorithm
Tagging a person	Manually	Automatically

II. SYSTEM ARCHITECTURE

As shown in Figure 1, the system consists of the following three components:

- User Database
- Face recognition Web service
- Friendship algorithm

A. User Database

In addition to the basic information of users, the information related to the face recognition feature is needed. The formation includes:

- Face - the user face identity, type and subject.
- Photo - photo information, ID and subject.
- Photocomment - the information of each user's comment on each photo.
- Friendship - the relationship information between each user.
- Recognition - the relationship information between the faces of photo and each friend.
- Invite - the relationship information of inviting people between user and friendship tables, inviting message and status of invitation.

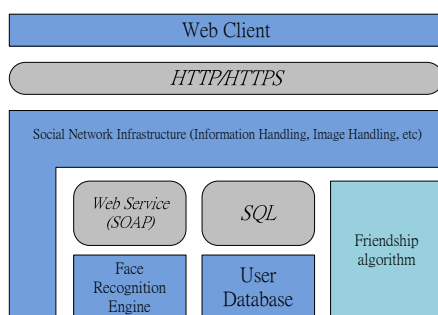


Figure 1. System architecture

B. Face Recognition Web Service

The face recognition web service handles the tasks involving face recognition. It implements with face recognition engine that face localization, enrollment and matching using digital image processing algorithm. The process of face recognition consists of a number of steps, which are shown in Figure 2.

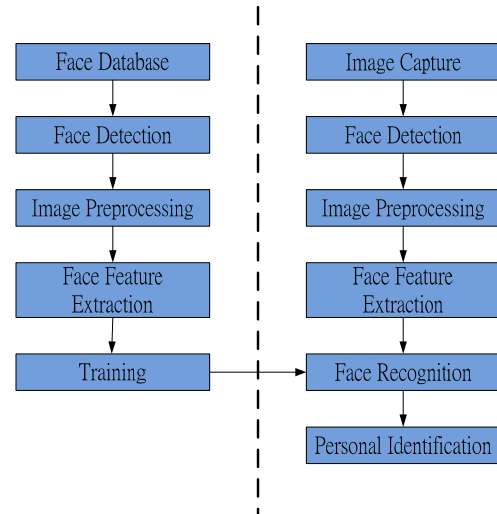


Figure 2. Common face recognition process

The normal procedures include two parts: the reference part and the identification part. In the reference process, the face recognition engine will first detect the faces for the raw image from the photo. It will then convert the photo to grayscale so as to eliminate the lighting effect and make some image process (PCA or ICA) to get a lower resolution from the face. After that, it will extract the face feature and store the face to the training database. In the face identification process, the engine will go through the same procedures to extract the face feature and make comparison with the training database. But in our implementation, we will do some normalization before store to the training database. We will normalize all face to the same resolution and size that can enhance

TABLE II.
CASES CONSIDERED IN THE FRIENDSHIP ALGORITHM

Case	Description	Unit	Weighting
My album	People appeared in my albums	No. of appearance	10
Indirect friend	Friend of my friend	No. of common friend	8
Friend's album	People appeared in my friend's albums	No. of appearance	6
Photo with me	Someone has photo that includes myself	No. of appearance	4
Co-photo	Someone and me in the same photo	No. of appearance	2

the speed of the identification face. After the face recognition step, it will provide a score for the face identity.

The face recognition service is provided by the way of Web services, which can make the system more scalability. When the system becomes larger, the face recognition process could become the performance bottleneck. Because the face recognition service is provided by web services, more face recognition engines can be added easily.

C. Friendship Algorithm

The friendship algorithm is used to provide the recommendation list for the users. The algorithm makes use the faces recognized in different cases to generate recommendation list. The cases used and the corresponding weighting for each case are shown in Table 2. Note that the exact values the weighting can be revised if necessary.

The friendship algorithm considers different case, and in each case there is a weighting associated. From the values of the weighting, it can be seen that priority has been considered in the friendship algorithm. That is, higher priority is given to the owner's direct relationship and lower priority to the indirect relationship.

As a result, a final score for each face can be generated by using the following formula:

$$\text{Final score} = \text{no. of appearance (my album)} \times 10 + \text{no. of common friend (indirect friend)} \times 8 + \text{no. of appearance (friend's album)} \times 6 + \text{no. of appearance (photo with me)} \times 4 + \text{no. of appearance (co-photo)} \times 2$$

The highest score of the face will appear the top of the recommend list, and the lower score faces will appear below the highest recommend friend according to the score.

Based on these final scores, the friendship algorithm will recommend the face to the user when the score higher than a threshold.

III. IMPLEMENTATION

A. Information Handling

The information handling process is the core development of the social network service, as shown in Figure 1. We used the PHP script language to develop this part. The functions are handling the user accounts, photo albums, searching and messaging. We use MySQL to implement our user database.

B. Face recognition process

During the face detection process, we make some optimizations:

- Convert image to grayscale.
- Extract the image with face detection.
- Resize the face to a suitable size.
- Store the face with base64-encoded string to face database.

The above optimizations can reduce the memory usage of each face and improve the performance of face matching. It can also reduce the process of storing the raw faces as the faces have been downsized. The based64-encoded string will be used in the implementation of the face reorganization web service.

Each user can store multiple face identities -- one main identity and other supplementary identities. So the user can store the current face as the main identity and some old faces as supplementary identities. These faces will store in the trained database for face matching uses.

After the comparison, the engine will give the face identity a score. We choose the score 50 as the basic recognition standard. Score 50 means nearly 0.01% of false acceptance rate (FAR). It is a very acceptable figure for the face recognition. Once the score of matching face is larger than 50, we would expect the matching face is the same as the reference face.

C. Face Recognition Engine

To implement the face recognition engine in our online social network, we use VeriLook SDK [8]. VeriLook SDK is a software development kit for face detection and face recognition. It supports multiplatform, Windows, Linux and MacOS. It also can support C/C++, C#, VB, Java and Delphi as development language. It contains two major components -- extractor and matcher. It also provides a camera manager library to support simultaneous capture from multiple cameras. The matching speed of the engine is 100,000 faces per second in 1:N identification. It supports live face detection, multiple face processing.

Our effort on this part is to make use the VeriLook SDK and to build a Web service for face recognition. We design this engine using Web service that can provide backend service with one or more servers. The advantage of making it a backend service is that if the face recognition technology changes in the future, the backend service can be replaced without the need to modify the other components.

Actually this prototype will only employ some simple procedures to detect and recognize faces. So the remote service is limited to five methods, DetectFaces, CompareFace, CompareMultipleFaces, CompareMultipleScore and CompareMultipleToMultipleFaces as shown in Figure 3. They covered all the necessary functions for face detection and recognition for our web service.

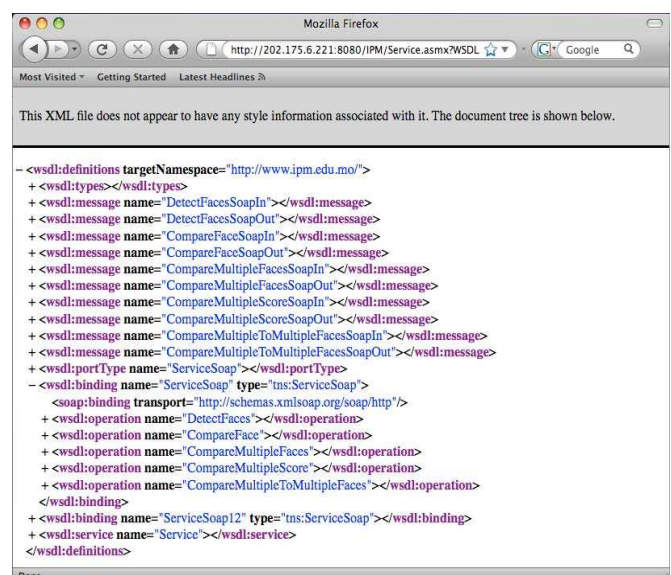


Figure 3. Web service definition

D. System Operation

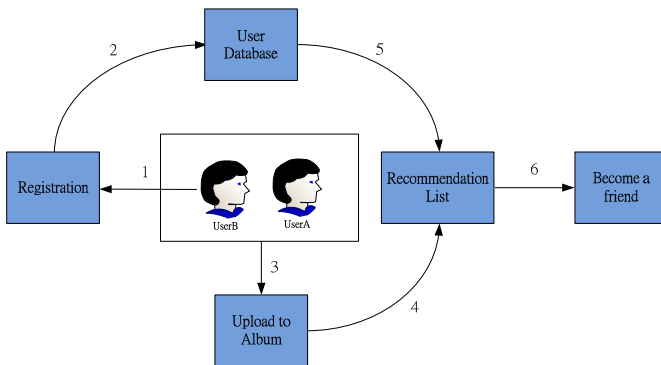


Figure 4. Establishing a friendship with the face recognition process

Figure 4 illustrate the system operation of how a user (UserB) meets another (UserA) using the face recognition technology.

1. Firstly, when UserB registers to the system, besides his basic personal information, he needs to submit a photo containing his face clearly.
2. When the system receives his face, it will look up the face in the database about UserB. If the face does not exist, it will add the face to the database.
3. Suppose UserA has taken a photo with UserB in a social event, and after the event, UserA uploaded the photo to his album in the system.
4. The system will try to detect the faces on the photos UserA just uploaded.
5. Since the system has the faces of UserA and UserB, it will put their relationship to the friendship algorithm to rank and make the recommendation lists for both UserA and UserB.
6. As a result, UserA can find UserB in his recommendation list, and vice versa. After the friend invitation by either UserA or UserB, they can become friend.

E. Testing

In the final stage of the development, a set of sample data was used for testing the functionality, acceptance and accuracy of the system. The sample data will include some text input, photo upload and comparison test. Some sample accounts and qualified photos are required for this test. We have collected the sample data of 30 users from existing social network application. The size of photo will be adjusted to around 30k for faster web page response and lower network latency. But the quality of the photos and faces must be maintained to an acceptable ratio.

The following testing procedures were run to ensure the correctness of the system:

- Create multiple user accounts.
- Upload the user photo that the face will be used as a face identity.
- Create albums with multiple accounts.
- Upload photos to multiple albums.
- Run the friendship algorithm scripts to create recommend lists.
- Make friends with or without the recommend lists.

- Click in the photo to examine the face tagging in album.

After testing the correct operations of the system, we have invited about 70 new users to use the system for testing user interface and performance. In the above testing stage, we have set up one machine for handling social network infrastructure and user database to serve web client, and set up another machine for handling face recognition engine to provide Web service. The configurations of the machines are:

- Core 2 Duo 3.0 GHz
- 1GB DDR2 Memory
- 250GB SATA II Hard Drive

IV. DEMONSTRATIONS

In this section, we will demonstrate how our online social network provides the features of recommendation list from photos, face identity database and face tagging.

A. Testing

Figure 5 shows the first page after the user logged in the system. The recommend list is shown on the lower left hand side. By clicking on the “Recognize my faces”, the user can upload photo to the face identity database. User may upload photo with the user face only or with other friends. The system will detect the faces and let the user choose the right face for the database. After this process, the user will have the face identity for the system. The user may have more than one face identity in the system.

When the user clicks the “edit” link beside the “Recognize my faces”, the corresponding face database will appear, as shown in Figure 6, so that the user can delete or choose the right head for the profile. This face database is actually a trained database of face recognition process. It stores all the possible face identities for face matching process in the friendship algorithm processing.

B. Face Tagging

Face tagging is a very useful feature in the system. The existing social network web sites commonly require a user to manually select and type name in the appropriate area of a picture. As our system employs the face recognition technology, it can recognize faces automatically without user interaction. If the user forgot the name of his friend, he can use the feature to list the name of his friend from the name tagging.



Figure 5. Web interface - my desktop



Figure 6. Web interface – face database

V. CONCLUSION

We implement a prototype to demonstrate the basic function of user registration, create and edit albums, upload and delete photos, edit some text fields, and make friends with/without recommend list.

Our online social network with face recognition technology provides features that improve user

experience and user friendly than the existing social networks. The features include recommendation list by photos, face identity database and face tagging.

ACKNOWLEDGMENT

The work described in this paper is supported by Macao Polytechnic Institute Research Grant (No.: RP/ESAP-1/2009).

REFERENCES

- [1] Barry, W. (1996). For a social network analysis of computer networks: a sociological perspective on collaborative work and virtual community.
- [2] W.W. Bledsoe, The model method in facial recognition, Panoramic Research, Inc., Palo Alto, CA PRI:15, August 1966.
- [3] M. Kirby, L. Sirovich, Application of the Karhunen-Loeve procedure for the characterization of human faces, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (1990) 103–107.
- [4] L. Sirovich, M. Kirby, A low-dimensional procedure for the characterization of human faces, Journal of the Optical Society of America 4 (1987) 519–524.
- [5] M. Turk, A. Pentland, Eigenfaces for recognition, Journal of Cognitive Neuroscience 3 (1991) 71–86.
- [6] M.S. Bartlett, H.M. Lades, T.J. Sejnowski, Independent component representations for face recognition, presented at SPIE Symposium on Electronic Imaging: Science and Technology, Conference on Human Vision and Electronic Imaging III, San Jose, CA, 1998.
- [7] M.S. Bartlett, J.R. Movellan, T.J. Sejnowski, Face recognition by independent component analysis, IEEE Transaction on Neural Networks 13 (2002) 1450–1464.
- [8] VeriLook SDK: <http://www.neurotechnology.com/>.

Ray K. C. Lai received the B.Sc degree in Computer Studies from Macao Polytechnic Institute, and is currently working toward the M.Sc degree in Software Engineering from University of Macau.

Jack C. K. Tang received the B.Sc degree in Computer Studies from Macao Polytechnic Institute. His research interests include online social networks and computer networks.

Angus K.Y. Wong is an associate professor at Macao Polytechnic Institute. His research interests include Internet systems, network infrastructure security, and human–computer interactions. Wong has a BSc and a PhD in information technology from the City University of Hong Kong.

Philip I.S. Lei is an associate professor at Macao Polytechnic Institute and a PhD candidate in computer science at Sun Yat-Sen University. His research interests include user interface design, social network analysis, and data mining.

Dynamic Differential Evolution for Constrained Real-Parameter Optimization

Youyun Ao¹, Hongqin Chi²

¹School of Computer and Information, Anqing Teachers College, Anqing, China

Email: youyun.ao@gmail.com

²Department of Computer, Shanghai Normal University, Shanghai, China

Email: chihq@shnu.edu.cn

Abstract—Differential evolution (DE) has been shown to be a simple and effective evolutionary algorithm for global optimization both in benchmark test functions and many real-world applications. This paper introduces a dynamic differential evolution (D-DE) algorithm to solve constrained optimization problems. In D-DE, a novel mutation operator is firstly designed to prevent premature. Secondly, the scale factor F and the crossover probability CR are dynamic and adaptive to be beneficial for adjusting control parameters during the evolutionary process, especially, when done without any user interaction. Thirdly, D-DE uses orthogonal design method to generate initial population and reinitialize some solutions to replace some worse solutions during the search process. Finally, D-DE is validated on 6 benchmark test functions provided by the CEC 2006 special session on constrained real-parameter optimization. The experimental results obtained by D-DE are explained and discussed, and some conclusions are also drawn.

Index Terms—constrained optimization, mutation scheme, differential evolution, evolutionary algorithm, constraint handling

I. INTRODUCTION

Many real-world optimization problems in science and engineering involve a number of constraints which the optimal solution must satisfy. These problems are also called constrained optimization problems or nonlinear programming problems. We are most interested in the general constrained optimization problems, which are all transformed into the following format [1], [2], [3], [4]:

$$\begin{aligned} &\text{Minimize } f(\vec{x}), \quad \vec{x} = [x_1, x_2, \dots, x_n] \in \mathbb{R}^n \\ &\text{Subject to } g_j(\vec{x}) \leq 0, \quad j = 1, 2, \dots, q \\ &\quad \quad \quad h_j(\vec{x}) = 0, \quad j = q + 1, q + 2, \dots, m \end{aligned} \quad (1)$$

Where $L_i \leq x_i \leq U_i, i = 1, 2, \dots, D$

Here n is the number of the decision or parameter variables (that is, \vec{x} is a vector of size D), the i th variable x_i varies in the range $[L_i, U_i]$. The function $f(\vec{x})$ is the objective function, $g_j(\vec{x})$ is the j th inequality constraint and $h_j(\vec{x})$ is the j th equality constraint. The decision or search space S is written as $S = \prod_{i=1}^D [L_i, U_i]$, and the feasible space F , expressed as $F = \{\vec{x} \in S \mid g_j(\vec{x}) \leq 0, j = 1, 2, \dots, q; h_j(\vec{x}) = 0, j = q + 1, q + 2, \dots, m\}$, is one subset

of the parameter space S (obviously, $F \subseteq S$) which satisfies the equality and inequality constraints.

Population-based evolutionary algorithm, mainly due to its ease to implement and use, and its less susceptibleness to the characteristics of the function to be optimized, has become a very popular option to solve constrained optimization problems in benchmark test functions and real-world applications [5]. Muñoz Zavala et al. [6] proposed a new constrained optimization algorithm based on improved particle swarm optimization (COPSO). In order to keep diversity, COPSO introduces a hybrid approach based on a modified ring neighborhood structure with two new perturbation operators for perturbing the particle swarm optimization (PSO) memory. In addition, COPSO adopts a new and special handling technique for equality constraints where a dynamic tolerance value is adjusted to allow the survival of unfeasible particles. Furthermore, COPSO is applied to the solution of state-of-the-art benchmark test functions and various engineering design problems. Liang and Suganthan [7] proposed a dynamic multi-swarm particle swarm optimizer with a novel constraint-handling mechanism (DMS-PSO). DMS-PSO adopts a novel constraint-handling mechanism based on multi-swarm. Different from the existing constraints handling methods, sub-swarms are adaptively assigned to explore different constraints during the search process. Additionally, DMS-PSO introduces Sequential Quadratic Programming (SQP) method to improve local search ability. Finally, DMS-PSO is applied to the solution of constrained real-parameter optimization. Mezura-Montes et al. [8] proposed a modified differential evolution for constrained optimization (MDE). In order to increase the probability of each parent to generate a better offspring, MDE allows each solution to generate more than one offspring but using a different mutation operator which combines information of the current parent to find new search directions. Besides, MDE employs three selection criteria based on feasibility to deal with the constraints and adopts a diversity mechanism to maintain infeasible solutions located in promising areas of the search space. Takahama and Sakai [9] proposed a novel constrained optimization algorithm by the ε constrained differential evolution with gradient-based mutation and feasible elites (ε DE). Firstly, ε DE applies the ε constrained method to differential evolution. Secondly, to solve problems with

many equality constraints faster, which are very difficult problems for numerical optimization, ε DE proposes gradient-based mutation and feasible elites preserving strategy. Finally, ε DE is tested on 24 benchmark test functions provided by the CEC 2006 special session on constrained real-parameter optimization. Differential evolution (DE) [10], [11], a relatively new evolutionary technique, has been shown to be simple and powerful and has been widely applied to both benchmark test functions and real-world applications [12]. After analyzing the existing evolutionary algorithms, this paper introduces a new dynamic differential evolution (D-DE) algorithm for constrained real-parameter optimization efficiently.

The remainder of this paper is organized as follows. Section II briefly introduces the basic idea of DE. Section III describes in detail the D-DE algorithm. Section IV presents 6 benchmark test functions. Section V presents experimental settings adopted by D-DE, conventional DE and GA, respectively. Section VI provides an analysis of the results obtained from our empirical study. Finally, some conclusions and some possible paths for future research are provided in Section VII.

II. THE BASIC IDEA OF CONVENTIONAL DE

Let us assume that $\vec{x}_i^t = [x_{i,1}^t, x_{i,2}^t, \dots, x_{i,D}^t]$ are solutions at generation t , $P^t = \{\vec{x}_1^t, \vec{x}_2^t, \dots, \vec{x}_N^t\}$ are population, where D denotes the dimension of solution space, N is population size. In conventional DE, the child population P^{t+1} is generated through following operators [10], [13]:

A. Mutation Operator

For each \vec{x}_i^t in parent population, the mutant vector \vec{v}_i^{t+1} is generated according to the following equation:

$$\vec{v}_i^{t+1} = \vec{x}_i^t + F \times (\vec{x}_{r_1}^t - \vec{x}_{r_2}^t) \quad (2)$$

Where $r_1, r_2, r_3 \in \{1, 2, \dots, N\} \setminus i$ are randomly chosen and mutually different, the scaling factor F is used to control amplification of the differential variation $\vec{x}_{r_1}^t - \vec{x}_{r_2}^t$.

B. Crossover Operator

For each individual \vec{x}_i^t , a trial vector \vec{u}_i^{t+1} is generated by the following equation:

$$u_{i,j}^{t+1} = \begin{cases} v_{i,j}^{t+1}, & \text{if } (rand \leq CR \parallel j = rand[1, D]) \\ x_{i,j}^t, & \text{otherwise} \end{cases} \quad (3)$$

Where $rand$ is a uniform random number distributed between 0 and 1, $rand[1, D]$ is a randomly selected index from the set $\{1, 2, \dots, D\}$, the crossover probability $CR \in [0, 1]$ is used to control the diversity of individuals.

C. Selection Operator

The child individual \vec{x}_i^{t+1} is selected from each pair of \vec{x}_i^t and \vec{u}_i^{t+1} by using greedy selection criterion:

$$\vec{x}_i^{t+1} = \begin{cases} \vec{u}_i^{t+1}, & \text{if } (f(\vec{u}_i^{t+1}) < f(\vec{x}_i^t)) \\ \vec{x}_i^t, & \text{otherwise} \end{cases} \quad (4)$$

Where the function f is the objective function and the condition $f(\vec{u}_i^{t+1}) < f(\vec{x}_i^t)$ means the individual \vec{u}_i^{t+1} is better than \vec{x}_i^t .

```

1: Generate initial population  $P^0$ .
2: Evaluate  $P^0$  and let generation counter  $t = 0$ .
3: While (the stopping criterion is not satisfied) do {
4:   For each individual  $\vec{x}_i^t$ , its offspring  $\vec{x}_i^{t+1}$  is generated
5:     by mutation, crossover and selection operators.
6: Evaluate  $P^{t+1}$  and let  $t = t + 1$  }
```

Figure 1. The general framework of DE.

III. THE PROPOSED ALGORITHM DDE

A. Orthogonal Initial Population

Generally, the initial population $P^0 = \{\vec{x}_1^0, \vec{x}_2^0, \dots, \vec{x}_N^0\}$ of evolutionary algorithm is randomly generated as follows:

$$\forall i \leq N, \forall j \leq D: x_{i,j}^0 = L_j + r_j \times (U_j - L_j) \quad (5)$$

Where N is the population size, D is the number of variables, r_j is a random number between 0 and 1, the j th variable of \vec{x}_i^0 , written as $x_{i,j}^0$, is initialized in the range $[L_j, U_j]$. In order to improve the search efficiency, this paper employs orthogonal design method to generate the initial population, which can make some points closer to the global optimal point and improve the diversity of solutions. The orthogonal design method is described as follows [14]:

For any given individual $\vec{x} = [x_1, x_2, \dots, x_D]$, the i th decision variable x_i varies in the range $[L_i, U_i]$. Here, each x_i is taken as each factor of orthogonal design. Let us assume that each factor holds Q levels, namely, quantize the domain $[L_i, U_i]$ into Q levels $\alpha_1, \alpha_2, \dots, \alpha_Q$.

The j th level of the i th factor $\alpha_{i,j}$ is defined as follows:

$$\alpha_{i,j} = \begin{cases} L_i, & j = 1 \\ L_i + (j-1) \left(\frac{U_i - L_i}{Q-1} \right), & 2 \leq j \leq Q-1 \\ U_i, & j = Q \end{cases} \quad (6)$$

Thereafter, we create the orthogonal array $M = (b_{i,j})_{N \times D}$ with D factors and Q levels, where N is the number of level combinations. The procedure of generating the orthogonal array $M = (b_{i,j})_{N \times D}$ is described as follows:

```

1: for ( $i = 1; i \leq N; i++$ )
2: {  $b_{i,1} = \text{int}((i-1)/Q) \bmod Q; b_{i,2} = (i-1) \bmod Q$  }
3: for ( $j = 3; j \leq D; j++$ )
4:   for ( $i = 1; i \leq N; i++$ )
5:     {  $b_{i,j} = (b_{i,1} \times (j-2) + b_{i,2}) \bmod Q$  }
6: Increment  $b_{i,j}$  by one for  $1 \leq i \leq N, 1 \leq j \leq D$ 
```

Figure 2. Generating orthogonal array $M = (b_{i,j})_{N \times D}$.

Therefore, the initial population $P^0 = (x_{i,j}^0)_{N \times D}$ can be generated by using the orthogonal array $M = (b_{i,j})_{N \times D}$, where the j th variable of individual \bar{x}_i^0 is $x_{i,j}^0 = a_{j,b_{i,j}}$.

B. Novel Mutation Scheme

According to the different variants of mutation, there are several different DE schemes often used, which are formulated as follows [10]:

$$\text{"DE/rand/1/bin": } \bar{v}_i^{t+1} = \bar{x}_{r_1}^t + F \times (\bar{x}_{r_2}^t - \bar{x}_{r_3}^t) \quad (7)$$

$$\text{"DE/best/1/bin": } \bar{v}_i^{t+1} = \bar{x}_{best}^t + F \times (\bar{x}_{r_1}^t - \bar{x}_{r_2}^t) \quad (8)$$

"DE/current to best/2/bin":

$$\bar{v}_i^{t+1} = \bar{x}_i^t + F \times (\bar{x}_{best}^t - \bar{x}_i^t) + F \times (\bar{x}_{r_1}^t - \bar{x}_{r_2}^t) \quad (9)$$

"DE/best/2/bin":

$$\bar{v}_i^{t+1} = \bar{x}_{best}^t + F \times (\bar{x}_{r_1}^t - \bar{x}_{r_2}^t) + F \times (\bar{x}_{r_3}^t - \bar{x}_{r_4}^t) \quad (10)$$

"DE/rand/2/bin":

$$\bar{v}_i^{t+1} = \bar{x}_{r_1}^t + F \times (\bar{x}_{r_2}^t - \bar{x}_{r_3}^t) + F \times (\bar{x}_{r_4}^t - \bar{x}_{r_5}^t) \quad (11)$$

Where \bar{x}_{best} is the best solution of the current population, and the control parameter F is usually set to be a constant. Using \bar{x}_{best} can improve the convergence speed but also increase the probability of getting stuck in the local optimum. In order to overcome the limitations, this paper proposes a novel variant of mutation, which is defined as follows:

$$\bar{v}_i^{t+1} = \bar{x}_{better} + F \times \sum_{k=1}^{K/2} (\bar{x}_{r_{2k-1}}^t - \bar{x}_{r_{2k}}^t) \quad (12)$$

Where $r_1, r_2, \dots, r_K \in \{1, 2, \dots, N\} \setminus i$, they are K mutually different and randomly chosen integers. The better solution \bar{x}_{better} is a random sample from top N_a solutions after ranking the current population based on the feasibility rule described in the later. The scale factor F is a dynamic control parameter and related to the generation number, which is defined as follows:

$$F = F_{\min} + (F_{\max} - F_{\min}) \times (1 - \frac{t}{T})^{F_b} \quad (13)$$

Here F_{\min} and F_{\max} are the bottom and upper boundaries of F , and usually are set to 0.1 and 0.9 respectively. The exponent F_b is a shape parameter determining the degree of dependency on the generation number and usually is set to 2 or 3. Two parameters t and T are the current generation number and the maximal generation number respectively.

C. Dynamic Control Parameters

In conventional DE, the crossover probability CR is a constant value between 0 and 1. In this study, a dynamic crossover probability CR is defined as follows:

$$CR = CR_{\min} + (CR_{\max} - CR_{\min}) \times (1 - \frac{t}{T})^{CR_b} \quad (14)$$

Here CR_{\min} and CR_{\max} are the bottom and upper boundaries of CR , and usually are set to 0.1 and 0.9 respectively. The exponent CR_b is a shape parameter

determining the degree of dependency on the generation number and usually is set to 2 or 3. Two parameters t and T are the current generation number and the maximal generation number respectively.

At the early stage, D-DE uses a bigger scale factor F and a bigger crossover probability CR to search the solution space to preserve the diversity of solutions and prevent premature; at the later stage, D-DE employs a smaller scale factor F and a smaller crossover probability CR to search the solution space to enhance the local search and prevent the better solutions found from being destroyed.

D. Repair Rule

After crossover, if one or more of the variables in the new vector \bar{u}_i^{t+1} are outside their boundaries, the violated variable value $\bar{u}_{i,j}^{t+1}$ is either reflected back from the violated boundary or set to the corresponding boundary value using the repair rule as follows [15], [16]:

$$\bar{u}_{i,j}^{t+1} = \begin{cases} (L_j + \bar{u}_{i,j}^{t+1})/2, & \text{if } (p \leq 1/3) \wedge (\bar{u}_{i,j}^{t+1} < L_j) \\ L_j, & \text{if } (1/3 < p \leq 2/3) \wedge (\bar{u}_{i,j}^{t+1} < L_j) \\ 2L_j - \bar{u}_{i,j}^{t+1}, & \text{if } (p > 2/3) \wedge (\bar{u}_{i,j}^{t+1} < L_j) \\ (U_j + \bar{u}_{i,j}^{t+1})/2, & \text{if } (p \leq 1/3) \wedge (\bar{u}_{i,j}^{t+1} > U_j) \\ U_j, & \text{if } (1/3 < p \leq 2/3) \wedge (\bar{u}_{i,j}^{t+1} > U_j) \\ 2U_j - \bar{u}_{i,j}^{t+1}, & \text{if } (p > 2/3) \wedge (\bar{u}_{i,j}^{t+1} > U_j) \end{cases} \quad (15)$$

Where p is a probability and uniformly distributed random number in the range $[0,1]$.

E. Constraint Handling Mechanism

In evolutionary algorithms for solving constrained optimization problems, the most common method to handle constraints is to use penalty functions. Usually equality constraints are transformed into inequalities of the form [4]:

$$|h_j(\bar{x})| - \varepsilon \leq 0, j = q+1, q+2, \dots, m \quad (16)$$

Here ε is a tolerance allowed (a very small value) for the equality constraints.

In general, the constraint violation function of one individual \bar{x} is transformed by m equality and inequality constraints as follows [9], [17], [18]:

$$G(\bar{x}) = \sum_{j=1}^q w_j \max(0, g_j(\bar{x}))^\beta + \sum_{j=q+1}^m w_j \max(0, |h_j(\bar{x})| - \varepsilon)^\beta \quad (17)$$

Here the exponent β is a positive number and usually set to 1 or 2, and the coefficient w_j is greater than zero. The function value $G(\bar{x})$ shows that the degree of constraints violation of individual \bar{x} . β is set to 2 and w_j is set to 1 in this study.

In this study, a simple and efficient constraint handling technique of feasibility-based rule is introduced, which is also a constraint handling technique without parameters. When two solutions are compared at a time, the following criteria are always applied [3]:

1) If one solution is feasible, and the other is infeasible, the feasible solution is preferred;

- 2) If both solutions are feasible, the one with the better objective function value is preferred;
- 3) If both solutions are infeasible, the one with smaller constraint violation function value is preferred.

F. Algorithm Framework

The general framework of the D-DE algorithm is outlined as follows:

```

1: Generate orthogonal initial population  $P^0 = \{\bar{x}_1^0, \bar{x}_2^0, \dots, \bar{x}_N^0\}$ ,
2: initialize parameters, and let  $t = 0$ .
3: Evaluate  $P^0$ , and rank  $P^0$  based on feasibility rule.
4: repeat
5:   for each individual  $\bar{x}_i^t$  in the population  $P^t$  do
6:     Generate  $K$  random integers  $r_1, r_2, \dots, r_K \in \{1, 2, \dots, N\} \setminus i$ ,
7:     they are also mutually different.
8:     Generate a random integer  $j_{rand} \in \{1, 2, \dots, D\}$ .
9:     Randomly select a sample  $\bar{x}_{better}$  from top  $N_a$  individuals of
10:    the current population  $P^t$ .
11:    for each parameter  $j \in \{1, 2, \dots, K/2\}$  do
12:      
$$\bar{u}_{i,j}^{t+1} = \begin{cases} \bar{x}_{better,j} + F \times (\bar{x}_{r_{2k-1},j}^t - \bar{x}_{r_{2k},j}^t), & \text{if } (rand \leq CR \parallel j = rand[1, D]) \\ \bar{x}_{i,j}^t, & \text{otherwise} \end{cases}$$

13:    end for
14:    Apply repair rule to repair  $\bar{u}_{i,j}^{t+1}$  if required, and evaluate  $\bar{u}_{i,j}^{t+1}$ .
15:    Replace  $\bar{x}_i^t$  with the child  $\bar{u}_i^{t+1}$  in the population  $P^{t+1}$ , if  $\bar{u}_i^{t+1}$ 
16:    is better, otherwise  $\bar{x}_i^t$  is retained.
17:  end for
18:  Rank  $P^{t+1}$  based on the feasibility rule, then replace  $N_b$  worse
19:  solutions with  $N_b$  orthogonal reinitialized solutions.
20:  Rank  $P^{t+1}$  based on the feasibility rule, and let  $t = t + 1$ .
21: until (the termination condition is achieved)

```

Figure 3. The general framework of the D-DE algorithm.

IV. TEST FUNCTION SUITE

In order to validate D-DE, we employ 6 benchmark test problems $g04$, $g06$, $g08$, $g11$, $g12$, which are provided by the CEC 2006 special session on constrained real-parameter optimization [4], and which are described in the following.

A. Test function $g02$

$$\text{Minimize } f(\bar{x}) = - \frac{\sum_{i=1}^n \cos^4(x_i) - 2 \prod_{i=1}^n \cos^2(x_i)}{\sqrt{\sum_{i=1}^n i x_i^2}}$$

$$\text{Subject to } g_1(\bar{x}) = 0.75 - \prod_{i=1}^n x_i \leq 0$$

$$g_2(\bar{x}) = \sum_{i=1}^n x_i - 7.5n \leq 0$$

Where $n = 20$ and $0 < x_i \leq 10$ ($i = 1, 2, \dots, n$). The global minimum $\bar{x}^* = (3.16246061572185, 3.12833142812967, 3.09479212988791, 3.06145059523469, 3.0279291588555, 2.99382606701730, 2.95866871765285, 2.92184227312450, 0.49482511456933, 0.48835711005490, 0.482316427$

11865, 0.47664475092742, 0.47129550835493, 0.46623099264167, 0.46142004984199, 0.45683664767217, 0.45245876903267, 0.44826762241853, 0.44424700958760, 0.44038285956317), the best is $f(\bar{x}^*) = -0.80361910412559$, constraint g_1 is close to being active.

B. Test function $g04$

$$\text{Minimize } f(\bar{x}) = 5.3578547x_3^2 + 0.8356891x_1x_5 + 37.293239x_1 - 40792.141$$

$$\text{Subject to } g_1(\bar{x}) = 85.334407 + 0.0056858x_2x_5 + 0.0006262x_1x_4 - 0.0022053x_3x_5 - 92 \leq 0$$

$$g_2(\bar{x}) = -85.334407 - 0.0056858x_2x_5 - 0.0006262x_1x_4 + 0.0022053x_3x_5 \leq 0$$

$$g_3(\bar{x}) = 80.51249 + 0.0071317x_2x_5 + 0.0029955x_1x_2 + 0.0021813x_3^2 - 110 \leq 0$$

$$g_4(\bar{x}) = -80.51249 - 0.0071317x_2x_5 - 0.0029955x_1x_2 - 0.0021813x_3^2 + 90 \leq 0$$

$$g_5(\bar{x}) = 9.300961 + 0.0047026x_3x_5 + 0.0012547x_1x_3 + 0.0019085x_3x_4 - 25 \leq 0$$

$$g_6(\bar{x}) = -9.300961 - 0.0047026x_3x_5 - 0.0012547x_1x_3 - 0.0019085x_3x_4 + 20 \leq 0$$

Where $78 \leq x_1 \leq 102$, $33 \leq x_2 \leq 45$ and $27 \leq x_i \leq 45$ ($i = 3, 4, 5$). The optimum solution is $\bar{x}^* = (78, 33, 29.9952560256815985, 45, 36.7758129057882073)$, where $f(\bar{x}^*) = -3.066553867178332e + 004$. Two constraints are active (g_1 and g_6).

C. Test function $g06$

$$\text{Minimize } f(\bar{x}) = (x_1 - 10)^3 + (x_2 - 20)^3$$

$$\text{Subject to } g_1(\bar{x}) = -(x_1 - 5)^2 - (x_2 - 5)^2 + 100 \leq 0$$

$$g_2(\bar{x}) = (x_1 - 6)^2 + (x_2 - 5)^2 - 82.81 \leq 0$$

Where $13 \leq x_1 \leq 100$ and $0 \leq x_2 \leq 100$. The optimum solution is $\bar{x}^* = (14.09500000000000064, 0.8429607892154795668)$ where $f(\bar{x}^*) = -6961.81387558015$. Both constraints are active.

D. Test function $g08$

$$\text{Minimize } f(\bar{x}) = - \frac{\sin^3(2\pi x_1) \sin(2\pi x_2)}{x_1^3(x_1 + x_2)}$$

$$\text{Subject to } g_1(\bar{x}) = x_1^2 - x_2 + 1 \leq 0$$

$$g_2(\bar{x}) = 1 - x_1 + (x_2 - 4)^2 \leq 0$$

Where $0 \leq x_1 \leq 10$ and $0 \leq x_2 \leq 10$. The optimum solution is $\bar{x}^* = (1.22797135260752599, 4.24537336612274885)$ where $f(\bar{x}^*) = -0.0958250414180359$.

E. Test function $g11$

$$\text{Minimize } f(\bar{x}) = x_1^2 + (x_2 - 1)^2$$

$$\text{Subject to } h(\bar{x}) = x_2 - x_1^2 = 0$$

Where $-1 \leq x_1 \leq 1$, $-1 \leq x_2 \leq 1$. The optimum solution is $\bar{x}^* = (-0.707036070037170616, 0.500000004333606807)$

where $f(\vec{x}^*) = 0.7499$.

F. Test function g12

Minimize

$$f(\vec{x}) = -(100 - (x_1 - 5)^2 - (x_2 - 5)^2 - (x_3 - 5)^2) / 100$$

$$\text{Subject to } g(\vec{x}) = (x_1 - p)^2 + (x_2 - q)^2 + (x_3 - r)^2 - 0.0625 \leq 0$$

Where $0 \leq x_i \leq 10 (i=1,2,3)$ and $p, q, r = 1, 2, \dots, 9$. The feasible region of the search space consists of 9^3 disjointed spheres. A point (x_1, x_2, x_3) is feasible if and only if there exists p, q, r such that the above inequality holds. The optimum solution is $\vec{x}^* = (5, 5, 5)$ where $f(\vec{x}^*) = -1$. The solution lies within the feasible region.

V. EXPERIMENTAL SETTINGS

A. Parameter Settings of D-DE

In our experimental study, the parameter values used in D-DE are set as follows: the population size $N = 50$, the maximal generation number $T = 5000$, the level number $Q = \lfloor \sqrt{N} \rfloor$, the number of top solutions $N_a = 0.1 \times N = 5$, the number of replaced solutions $N_b = 0.1 \times N = 5$, the minimal and maximal values of scale factor F are set to $F_{\min} = 0.1$ and $F_{\max} = 0.9$ respectively, $K = 6$, the minimal and maximal values of crossover probability CR are set to $CR_{\min} = 0.1$ and $CR_{\max} = 0.9$ respectively, the shape parameter values $F_a = 3$ and $F_b = 3$ respectively, the exponent $\beta = 2$, the tolerant value $\varepsilon = 0.0001$. The number of function evaluations (FES) is equal to $(1 + N_b) \times N \times T = 275,000$. The achieved solution at the end of $(1 + N_a) \times N \times T$ FES is used to measure the performance of D-DE. D-DE is independently run 30 times on each test function.

B. Parameter Settings of Conventional DE

In our experimental study, the parameter values adopted by the conventional DE are set as follows: the population size $N = 50$, the maximal generation number $T = 55000$, the crossover probability $CR = 0.9$, the scale factor $F = 0.6$, the tolerant value $\varepsilon = 0.0001$. The number of function evaluations (FES) is equal to $N \times T = 275,000$. The achieved solution at the end of $N \times T$ FES is used to measure the performance of the conventional DE. The DE employs the repair rule and constraint handling mechanism described in Section III and is independently run 30 times on each test function.

C. Parameter Settings of Conventional GA

As a computing technique and method, population-based genetic algorithm (GA) [20] has been shown to be an effective evolutionary algorithm [21]. In our experimental study, the conventional GA uses simulated binary crossover (SBX) [22], polynomial mutation operator [23], tournament selection between the parent and its child, the repair rule and constraint handling mechanism described

in Section III. The parameter values employed by the real-coded GA are set as follows: the population size $N = 50$, the maximal generation number $T = 55000$, the crossover probability $P_c = 0.9$ and a mutation probability $P_m = 1/n$ (where n is the number of decision variables for real-coded GA), the distribution indexes for crossover and mutation operators as $\eta_c = 20$ and $\eta_m = 20$; the tolerant value $\varepsilon = 0.0001$. The number of function evaluations (FES) is equal to $N \times T = 275,000$. The obtained solution at the end of $N \times T$ FES is used to measure the performance of the conventional GA. The GA is independently run 30 times on each test function.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Comparison with Respect to Conventional DE and GA on 6 Benchmark Test Problems

Using the above experimental settings, the best, mean and worst results obtained by D-DE, DE, and GA are given in Tables I-III. As shown in Table I, D-DE, DE and GA all can find the best solution for each test problems g08, g11 and g12, respectively. For test problems g04 and g06, D-DE and DE can find the best solution and is better than that of GA. For test problem g02, the best result obtained by DE is slightly better than that of D-DE, and obviously better than that of GA. According to the mean results given in Table II, the mean result obtained by D-DE is better than or not worse than that of DE and GA for all test problems g02, g04, g06, g08, g11 and g12, while the mean result obtained by DE is better than or not worse than that of GA for test problems g02, g04, g06, g08 and g12 except for test problem g11 where the mean result obtained by GA is better than that of DE. Table II shows that the worst result obtained by D-DE is better than or not worse than that of DE and GA for all test problems g02, g04, g06, g08, g11 and g12, while the worst result obtained by DE is better than or not worse than that of GA for test problems g02, g04, g06, g08 and g12, except for test problem g11 where the worst result obtained by GA is better than that of DE. Additionally, according to Tables I-III, we can find that D-DE can almost find the optimum solution for each test problems g02, g04, g06, g08, g11 and g12 in one single run, and that D-DE is robust and can outperform DE and GA on a set of test problems.

TABLE I.
THE BEST RESULTS OBTAINED BY D-DE WITH RESPECT TO THOSE
OBTAINED BY DE, GA ON 6 BENCHMARK PROBLEMS.

Function	Optimal	D-DE	DE	GA
g02	-0.803619	-0.80356676	-0.80361902	-0.80254846
g04	-30665.539	-30665.53867	-30665.53867	-30664.86146
g06	-6961.814	-6961.81388	-6961.81388	-6958.24296
g08	-0.095825	-0.09582504	-0.09582504	-0.09582504
g11	0.7499	0.749900000	0.749900000	0.749900232
g12	-1	-1	-1	-1.00000000

TABLE II.
THE MEAN RESULTS OBTAINED BY D-DE WITH RESPECT TO THOSE
OBTAINED BY DE, GA ON 6 BENCHMARK PROBLEMS.

Function	Optimal	D-DE	DE	GA
g02	-0.803619	-0.80150890	-0.79516035	-0.78941723
g04	-30665.539	-30665.53867	-30665.53867	-30661.64864
g06	-6961.814	-6961.81388	-6961.81388	-6925.57354
g08	-0.095825	-0.09582504	-0.09582504	-0.09582504
g11	0.7499	0.749900000	0.897159582	0.750805569
g12	-1	-1	-1	-1.00000000

TABLE III.
THE WORST RESULTS OBTAINED BY D-DE WITH RESPECT TO THOSE
OBTAINED BY DE, GA ON 6 BENCHMARK PROBLEMS.

Function	Optimal	D-DE	DE	GA
g02	-0.803619	-0.79253455	-0.77041284	-0.75131833
g04	-30665.539	-30665.53867	-30665.53867	-30655.21039
g06	-6961.814	-6961.81388	-6961.81388	-6887.72373
g08	-0.095825	-0.09582504	-0.09582504	-0.09582504
g11	0.7499	0.749900000	1.000000000	0.755040572
g12	-1	-1	-1	-1.00000000

B. Comparison with Respect to Some State-of-the-art Approaches on 6 Benchmark Test Problems

In this section, we present the experimental results in detail and compare D-DE with respect to state-of-the-art algorithms. The experimental results are given in Table IV. The optimized objective function values (of 30 runs) arranged in ascending order and the 15th value in the list

is called the median optimized function value.

According to the summary of statistical results of test problems g04, g06, g08, g11, g12 given in Table IV, it is clearly seen that D-DE, A-DDE [19], COPSO [6] and SRES [18] all can find the optimum or near-optimum, when D-DE uses 275,000 FES, A-DDE 180,000 FES, COPSO 350,000 FES and SRES 500,000 FES, respectively. For test problem g02, the mean, worst and standard derivation of values obtained by D-DE are the best when compared with other algorithms, while the median value obtained by D-DE is better than that of A-DDE and SRES, and is slightly worse than that of COPSO. Besides, for test problem g02, the best value obtained by D-DE is slightly worse than that of the other algorithms. As shown in Table V, the best, median, mean, worst and standard derivation of values obtained by D-DE when set to 550,000 FES are obviously better than those when set to 275,000 FES. The best, median values obtained by D-DE when set to 550,000 FES are almost convergent to the optimum or near-optimum. Therefore, for test problem g02, D-DE is not still convergent to the optimum when set to 275,000 FES. In conclusion, the performance of D-DE is stable and better than or not worse than some state-of-the-art evolutionary algorithms on a set of test problems.

TABLE IV.
COMPARISON D-DE WITH RESPECT TO ALGORITHMS A-DDE [19], COPSO [6], SRES [18] ON 6 BENCHMARK TEST FUNCTIONS

Function	Optimal	Method	Best	Median	Mean	Worst	Std	FES
g02	-0.803619	D-DE	-0.80356676178	-0.803457738386	-0.801508902296	-0.79253454688	4.00E-03	275,000
		A-DDE	-0.803605	-0.777368	-0.771090	-0.609853	3.66E-02	180,000
		COPSO	-0.803619	-0.803617	-0.801320	-0.786566	4.59E-03	350,000
		SRES	-0.804	-0.793	-0.788	-0.746	1.3E-02	500,000
g04	-30665.539	D-DE	-30665.53867178	-30665.53867178	-30665.53867178	-30665.53867178	1.16E-011	275,000
		A-DDE	-30665.539	-30665.539	-30665.539	-30665.539	3.20E-13	180,000
		COPSO	-30665.538672	-30665.538672	-30665.538672	-30665.538672	0	350,000
		SRES	-30665.539	-30665.539	-30665.539	-30665.539	0.0E+00	500,000
g06	-6961.814	D-DE	-6961.81387558	-6961.81387558	-6961.81387558	-6961.81387558	4.63E-012	275,000
		A-DDE	-6961.814	-6961.814	-6961.814	-6961.814	2.11E-12	180,000
		COPSO	-6961.813876	-6961.813876	-6961.813876	-6961.813876	0	350,000
		SRES	-6961.814	-6961.814	-6961.814	-6961.814	1.9E-12	500,000
g08	-0.095825	D-DE	-0.095825041418	-0.095825041418	-0.095825041418	-0.095825041418	2.82E-017	275,000
		A-DDE	-0.095825	-0.095825	-0.095825	-0.095825	9.10E-10	180,000
		COPSO	-0.095825	-0.095825	-0.095825	-0.095825	0	350,000
		SRES	-0.096	-0.096	-0.096	-0.096	0.0E+00	500,000
g11	0.7499	D-DE	0.749900000000	0.749900000000	0.749900000000	0.749900000000	1.13E-016	275,000
		A-DDE	0.75	0.75	0.75	0.75	5.35E-15	180,000
		COPSO	0.749999	0.749999	0.749999	0.749999	0	350,000
		SRES	0.750	0.750	0.750	0.750	1.1E-16	500,000
g12	-1	D-DE	-1	-1	-1	-1	0	275,000
		A-DDE	-1.000	-1.000	-1.000	-1.000	4.10E-11	180,000
		COPSO	-1.000000	-1.000000	-1.000000	-1.000000	0	350,000
		SRES	-1.000	-1.000	-1.000	-1.000	0.0E+00	500,000

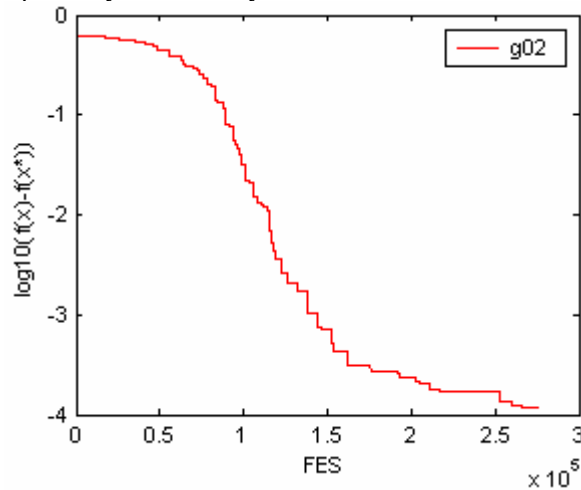
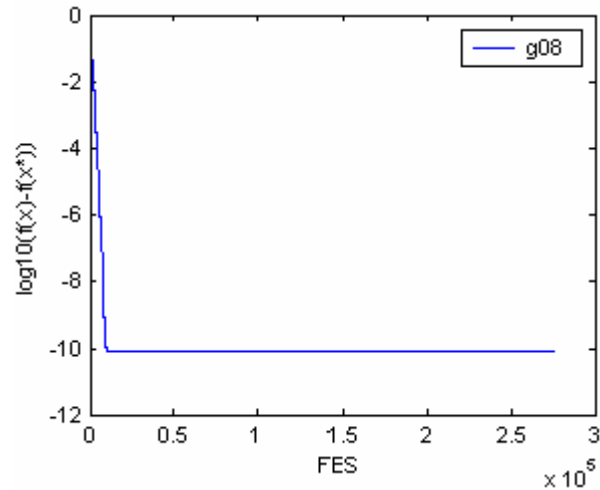
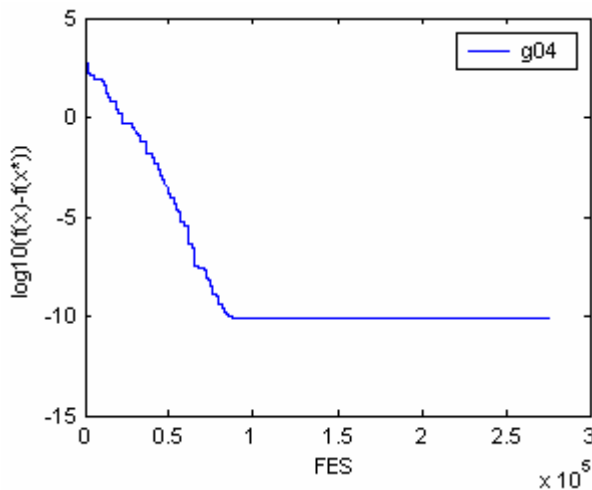
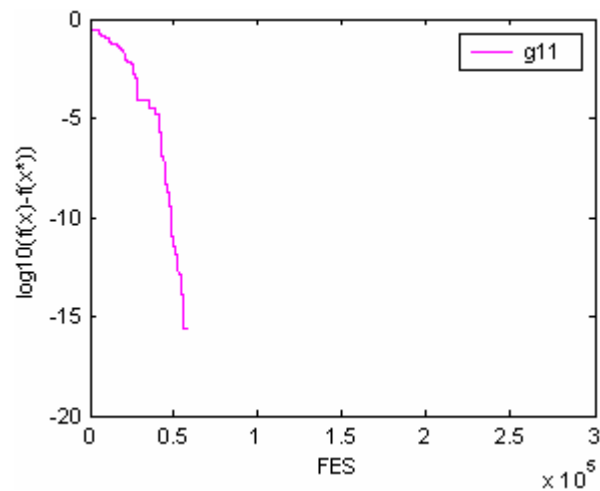
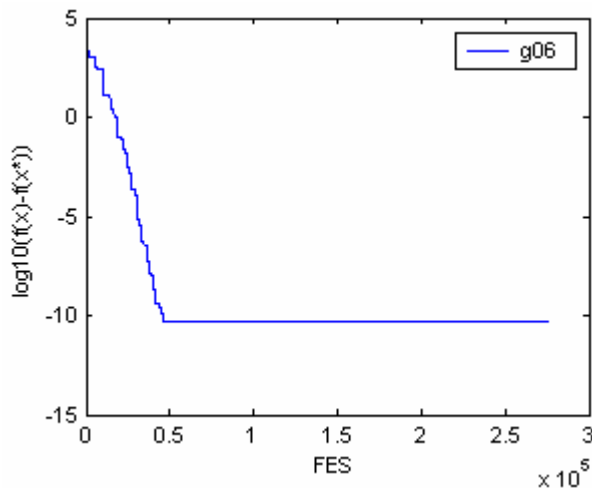
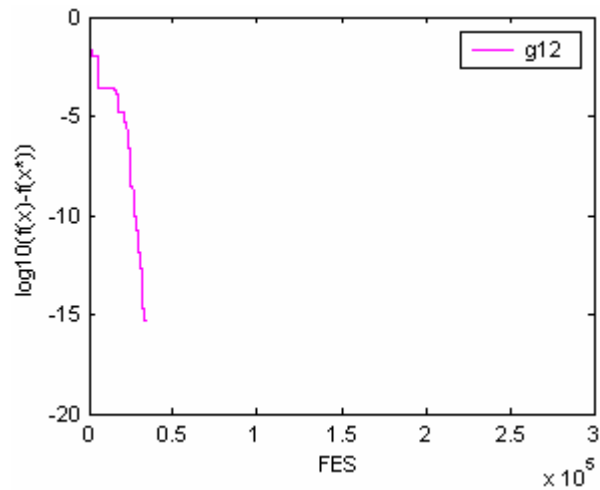
TABLE V.
EXPERIMENTAL RESULTS OBTAINED BY D-DE WHEN FES=275,000, FES=550,000 FOR TEST FUNCTION g02 OVER 30 RUNS

FES	Best	Median	Mean	Worst	Std
275,000	-0.80356676178	-0.803457738386	-0.801508902296	-0.79253454688	4.00E-03
550,000	-0.803610090279	-0.8036058890	-0.802935868229	-0.79259834414	2.55E-03

C. Convergence Graphs Obtained by D-DE for 6 Benchmark Test Problems

In order to provide a more intuitive comprehension, we present the convergence graphs obtained by D-DE for test problems $g02$, $g04$, $g06$, $g08$, $g11$ and $g12$. Figures 4-9 depict the convergence graphs for test problems $g02$, $g04$, $g06$, $g08$, $g11$ and $g12$, respectively. It is clearly seen that D-DE has a trend to

find the optimum solution for test problem $g02$ within 300,000 FES, that D-DE can find the optimum solution for each test problem $g06$, $g08$, $g11$, $g12$ within 50,000 FES, and that D-DE can obtain the optimum solution for test problem $g04$ within 100,000 FES.

Figure 4. Convergence graph for $g02$.Figure 7. Convergence graph for $g08$.Figure 5. Convergence graph for $g04$.Figure 8. Convergence graph for $g11$.Figure 6. Convergence graph for $g06$.Figure 9. Convergence graph for $g12$.

VII. CONCLUSIONS AND FUTURE WORK

In this study, we present a dynamic differential evolution algorithm (D-DE) for solving constrained real-parameter optimization problems. In this model of D-DE, there exist at least three important contributions as follows:

1) The first contribution is the novel mutation scheme, which can improve the convergence speed, prevent premature and preserve the diversity of solutions.

2) The second contribution is two important control parameters (i.e., the scale factor F and the crossover probability CR), which are dynamic and beneficial for adjusting control parameters during the evolutionary and search process, especially, when done without any user interaction.

3) The third contribution is that D-DE can prevent premature and enhance the search performance mainly due to replacing some relatively worse solutions with reinitialized solutions during the evolutionary process.

In addition, D-DE employs orthogonal design method to generate initial population to improve the diversity of solutions and introduces a constraint handling technique based on the feasibility rule and the sum of constraints violation.

Finally, D-DE is tested on 6 benchmark test functions provided by the CEC 2006 special session on constrained real-parameter optimization. Through comparing D-DE with respect to state-of-the-art evolutionary algorithms, the experimental results show that D-DE is highly competitive and can obtain good results in terms of a test set of constrained real-parameter optimization problems. However, in the future, there are still many aspects to do. Firstly, in order to further validate D-DE, we are considering of the possibility of testing more benchmark test functions (especially, highly dimensional problems) and real-world constrained optimization problems. Secondly, for some test functions, there exists the phenomenon of slow evolutionary at the later stage. In order to overcome the limitation, we will incorporate some local search techniques into D-DE to improve the convergence speed. Additionally, improving constraint handling technique is another future work.

REFERENCES

- [1] V. L. Huang, A. K. Qin, and P. N. Suganthan, "Self-adaptive differential evolution algorithm for constrained real-parameter optimization," in *2006 IEEE Congress on Evolutionary Computation (CEC'2006)*, pp. 324-331, IEEE, Vancouver, BC, Canada, July 2006.
- [2] A. E. Muñoz-Zavala, A. Hernández-Aguirre, E. R. Villadharce, and S. Botello-Rionda, "PESO+ for Constrained Optimization," in *2006 IEEE Congress on Evolutionary Computation (CEC'2006)*, pp. 935-942, IEEE, Vancouver, BC, Canada, July 2006.
- [3] K. Deb, "An efficient constraint handling method for genetic algorithms," *Computer Methods in Applied Mechanics and Engineering*, Vol. 186, No. 2, pp. 311-338, 2000.
- [4] J. J. Liang, T. P. Runarsson, E. Mezura-Montes, M. Clerc, P. N. Suganthan, C. A. Coello Coello, and K. Deb, "Problem definitions and evaluation criteria for the CEC 2006 special session on constrained real-parameter optimization," *Technical Report*, Nanyang Technological University, Singapore, 2006.
- [5] R. Landa-Becerra, C. A. Coello Coello, "Cultured differential evolution for constrained optimization," *Computer Methods in Applied Mechanics and Engineering*, Vol. 195, No. 33-36, pp. 4303-4322, 2006.
- [6] A. E. Muñoz Zavala, A. Hernández Aguirre, E. R. Villa Diharce, and S. Botello Rionda, "Constrained optimization with an improved particle swarm optimization algorithm," *International Journal of Intelligent Computing and Cybernetics*, Vol. 1, No. 3, pp. 425-453, 2008.
- [7] J. J. Liang, P. N. Suganthan, "Dynamic multi-swarm particle swarm optimizer with a novel constraint-handling mechanism," in *2006 IEEE Congress on Evolutionary Computation (CEC'2006)*, pp. 316-323, IEEE, Vancouver, BC, Canada, July 2006.
- [8] E. Mezura-Montes, J. Velázquez-Reyes, and C. A. Coello Coello, "Modified differential evolution for constrained optimization," in *2006 IEEE Congress on Evolutionary Computation (CEC'2006)*, pp. 332-339, IEEE, Vancouver, BC, Canada, July 2006.
- [9] T. Takahama, and S. Sakai, "Constrained optimization by the ϵ constrained differential evolution with gradient-based mutation and feasible elites," in *2006 IEEE Congress on Evolutionary Computation (CEC'2006)*, pp. 308-315, IEEE, Vancouver, BC, Canada, July 2006.
- [10] R. Storn, K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, Vol. 11, pp. 341-359, 1997.
- [11] K. Price, R. Storn, J. Lampinen, *Differential Evolution: A Practical Approach To Global Optimization*, Berlin: Springer-Verlag, 2005.
- [12] Z. Y. Yang, K. Tang, X. Yao, "Self-adaptive differential evolution with neighborhood search," *2008 Congress on Evolutionary Computation (CEC'2008)*, pp. 1110-1116, 2008.
- [13] H. A. Abbass, R. Sarker, C. Newton, "PDE: a Pareto-frontier differential evolution approach for multiobjective optimization problems," in *Proceedings of IEEE Congress on Evolutionary Computation*, Vol. 2, pp. 971-978, 2001.
- [14] Y. W. Leung, Y. P. Wang, "An orthogonal genetic algorithm with quantization for global numerical optimization," *IEEE Transactions on Evolutionary Computation*, Vol. 5, No. 1, pp. 40-53, 2001.
- [15] J. Brest, V. Zumer, and M. S. Maucec, "Self-adaptive differential evolution algorithm in constrained real-parameter optimization," in *2006 IEEE Congress on Evolutionary Computation (CEC'2006)*, pp. 919-926, IEEE, Vancouver, BC, Canada, July 2006.
- [16] Y. Wang, Z.X. Cai, "A Hybrid Multi-Swarm Particle Swarm Optimization to Solve Constrained Optimization Problems," *Frontiers of Computer Science in China*, Vol. 3, No. 1, pp. 38-52, 2009.
- [17] C. A. Coello Coello, "Theoretical and Numerical Constraint-Handling Techniques used with Evolutionary Algorithms: A Survey of the State of the Art," *Computer Methods in Applied Mechanics and Engineering*, Vol. 191, No. 11-12, pp. 1245-1287, 2002.
- [18] T. P. Runarsson, "Approximate evolution strategy using stochastic ranking," in *2006 IEEE Congress on Evolutionary Computation (CEC'2006)*, pp. 2760-2767, IEEE, Vancouver, BC, Canada, July 2006.
- [19] E. Mezura-Montes, A. G. Palomeque-Ortiz, "Parameter control in differential evolution for constrained optimization," in *2009 IEEE Congress on Evolutionary Computation (CEC'2009)*, pp. 1375-1382, 2009.

- [20] Z. Michalewicz, *Genetic Algorithms + Data Structures = Evolution Programs*, Springer-Verlag, third edition, 1996.
- [21] E. Mezura-Montes, C. A. Coello Coello, "A simple multimembered evolution strategy to solve constrained optimization problems," *IEEE Transactions on Evolutionary Computation*, Vol. 9, No. 1, pp. 1-17, 2005.
- [22] K. Deb and R. B. Agrawal, "Simulated binary crossover for continuous search space," *Complex Systems*, Vol. 9, No. 2, pp. 115-148, 1995.
- [23] K. Deb and M. Goyal, "A robust optimization procedure for mechanical component design based on genetic

adaptive search," *Transactions of the ASME: Journal of Mechanical Design*, Vol. 120, No. 2, pp. 162-164, 1998.

Youyun Ao was born in 1973. He received his B.S. degree in Computer Science from Jiangxi Normal University, Nanchang, China in 1999. He received his M.S. degree in Computer Software and Theory from Shanghai Normal University, Shanghai, China in 2006. He is currently a lecturer in Computer Science at Anqing Teachers College, Anqing, Anhui, China. His research interests include evolutionary computation, intelligent information processing, etc.

Fuzzy Logic Based Position-Sensorless Speed Control of Multi Level Inverter Fed PMBLDC Drive

Narmadha T.V.

Research Scholar, Anna University, Chennai, India
nar_velu@yahoo.co.in

Thyagarajan T.

Prof. & Head, Dept. of Instrumentation Engg,
 Anna University, Chennai, India
thyagu_vel@yahoo.co.in

Abstract—This paper presents multi level inverter fed Permanent Magnet Brushless DC Motor (PMBLDCM) with a simplified voltage control technique. It is based on the “indirect position sensing,” which was justified by the observation that position sensing came indirectly from voltage and current waveforms. The switching angle for the pulse is selected in such way to reduce the harmonic distortion. This drive system has advantages like reduced total harmonic distortion and higher torques. PI, Fuzzy and Hybrid (Fuzzy and PI) controllers are discussed. Closed loop simulation response is obtained for PI, Fuzzy and Hybrid controller with a disturbance in the input source. The conventional circuit is improved by introducing Hybrid Controller. In Industrial application the physical integration of Hybrid controller in the motor body itself is able to make them most suitable for low power (0.5hp) blowers and low power (50W) tube axial fans for cooling the electronic equipment. The performance of the PMBLDCM system is simulated and implemented. Simulation results of these systems are presented and the performance measures are compared. The simulation results with Hybrid controller indicate improved performance. The experimental results are compared with simulation results.

Index Terms : Three level inverter, FLC, PI, Hybrid, trapezoidal back emf, PWM, , Sim Power systems.

I. INTRODUCTION

With the rapid development of microelectronics and power switches, most adjustable-speed drives are now realized with ac machines. Permanent Magnet Synchronous Motor (PMSM) with sinusoidal shape back-EMF and brushless DC (BLDC) motor with trapezoidal shape back-EMF drives have been extensively used in many applications, ranging from servo to traction drives due to several distinct advantages such as high power density, high efficiency, large torque to inertia ratio, and better controllability. Brushless DC motor (BLDC) fed by two-phase conduction scheme has higher power/weight, torque/current ratios and it is less expensive due to the concentrated windings which shorten the end windings compared to three-phase permanent magnet synchronous motor (PMSM) [1]- [6]. There are two methods of

controlling PMBLDC motor namely sensor control and sensorless control. The latter has advantages like cost reduction, reliability, elimination of difficulty in maintaining the sensor etc. Sensorless control is highly advantageous when the motor is operated in dusty or oily environment, where cleaning and maintaining of Hall Sensors is required for proper sensing of rotor position. Sensorless method is preferred when the motor is in less accessible location. Accommodation of position sensor in motor used in compact unit such as computer hard disk may not be possible. Novel direct back emf detection for sensorless BLDC motor is given in [7]. Analysis of BLDC motor is given in [8]. Modeling of BLDC motor is given in [9]. Feed forward speed control of Brushless DC motor with input shaping is given in [10]. A PSO-based optimization of PID controller for a Linear BLDC Motor is given in [11]. Speed Control of BLDC based on CMAC & PID controller is given in [12]. A sensorless drive system for BLDC using a Digital Phase-Locked Loop is given in [13]. Classical control methods can be implemented in well- defined systems to achieve good performance of the systems. To control a system, an accurate mathematical model of the complete system is required. Systems with non linear behavior cannot be exactly modeled. The fuzzy Logic control has adaptive characteristics that can achieve robust response to a system with uncertainty, parameter variations and load disturbance. Fuzzy Logic and Fuzzy set theory was presented by Zadeh [14]. Fuzzy Logic Controllers have been broadly used for ill-defined, non-linear and complex systems [15], [16]. In the area of electrical drives, fuzzy logic controllers have been applied to switched reluctance motors [17], [18], induction motors [19] and PMBLDC motors [20] successfully. The above literature does not deal with voltage control method to control the speed using sensorless approach. This paper demonstrates a sensorless technique to drive a three phase brushless DC motor with a multi level voltage Inverter system using voltage control method with Hybrid Fuzzy logic control. PMBLDC motors drives are used in a wide range of commercial and residential applications such as domestic appliances, heating, ventilating and air-conditioning

Manuscript received June 24, 2009; revised August 31, 2009.

equipment due to their highest possible efficiencies. The speed control ability using Hybrid fuzzy controller is able to provide operation at their high efficiency.

In Section I Voltage control based PMBLDC motor is described. In Section II three level inverter is presented with some basics of mode of conduction. In Section III three level inverter fed PMBLDC motor with PI controller is presented. In Section IV three level inverter fed PMBLDC motor with Fuzzy Logic controller is presented. Hardware circuit is fabricated and tested.

A. VOLTAGE CONTROL BASED PMBLDC MOTOR

The measurement of armature current in the three phases is not required because there is no neutral connection, and hence the third phase current can be obtained from the other two. The quasi square-wave armatures current are mainly characterized through their maximum amplitude value, which directly controls the machine torque. The inverter performance is very much reliable because there is natural dead time for each transistor. Hence, allows designing a circuit for controlling only a DC component, which represents the maximum amplitude value of the trapezoidal currents, I_{MAX} , and reduces the complex circuitry required by other machines, allowing the self-synchronization process for the operation of the machine. The most popular way to control BLDCM for traction applications is through voltage-source current-controlled inverters. The inverter must supply a quasi-square current waveform whose magnitude, I_{MAX} , is proportional to the machine shaft torque. Then, by controlling the phase-currents, torque and speed can be adjusted. The response of phase voltage, phase currents, speed, and feedback current with disturbance at the source are obtained.

The waveforms of the armature currents are quasi-square. These currents are sensed through current sensors. These signals are then rectified, and a DC component, with the value of the ceiling of the currents, I_{max} , is obtained. It is filtered and DC voltage is obtained across resistor. The voltage V_{max} is compared with V_{ref} and from this comparison, an error signal "e(t)" is obtained. This error signal is then processed using PI controller. The output of PI controller is used to vary the input voltage of three level inverter. The strategy becomes simple, because the control only needs to be in command of DC current instead of three alternating waveforms. The control strategy also allows regenerative braking, which is very important in applications, like electric vehicles, where energy can be returned to the battery pack.

II. THREE LEVEL INVERTER

In three level inverter modeling, 120 degree conduction mode is employed. The gating signals given to the MOSFET are sequenced to every 60 degree interval. Each MOSFET conducts for a duration of 120 degrees. The MOSFET is used as a switch since it can operate high switching frequency. This feature is helpful in driving the

motor with high current and low voltage conditions. The three level inverter circuit is shown in Fig 1.

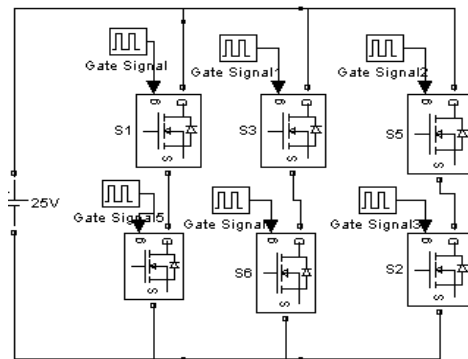


Fig 1. Three level Inverter Circuit

A. THREE LEVEL INVERTER FED PMBLDC MOTOR

Fig. 2 shows the schematic diagram of the Closed loop Sensorless Speed Control of PMBLDC motor using PI Controller. The MOSFETs are used as switching devices. For speed control of motor, the output frequency of the inverter is varied. The applied voltage to the motor is varied in linear proportion to the supply frequency to maintain the flux constant. The MATLAB simulation is carried out and the simulation results are presented in this section. The Fig.3 shows the driving pulses applied to the MOSFETs.

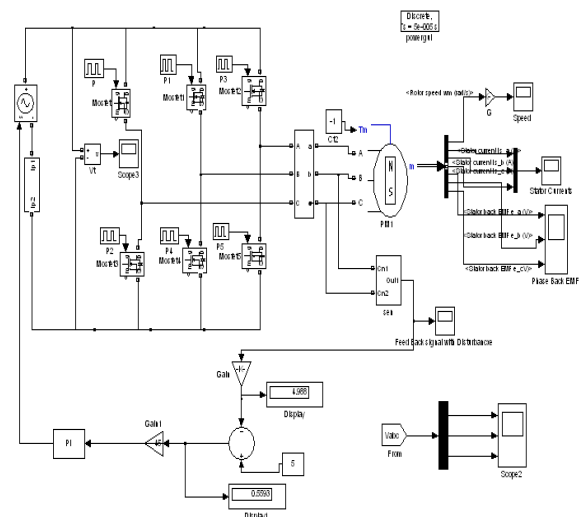


Fig 2. Closed loop Sensorless Speed Control of PMBLDC motor

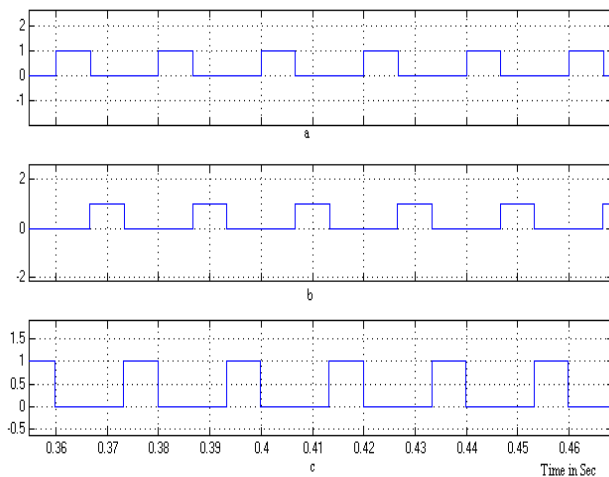


Fig3 a,b,c.Driving Pulses to the MOSFETs

B.THREE LEVEL INVERTER FED PMBLDC MOTOR WITH FUZZY LOGIC CONTROLLER

Fuzzy Logic Controller

The block diagram showing the implementation of the Fuzzy speed controller is illustrated in Figure 4. It includes four major blocks: knowledge base, fuzzification, inference mechanism, and defuzzification. The knowledge base is composed of a data and a rule base. The data base, consisting of input and output membership functions. The rule base is made of a set of linguistic rules relating the fuzzy input variables into the desired fuzzy control actions.

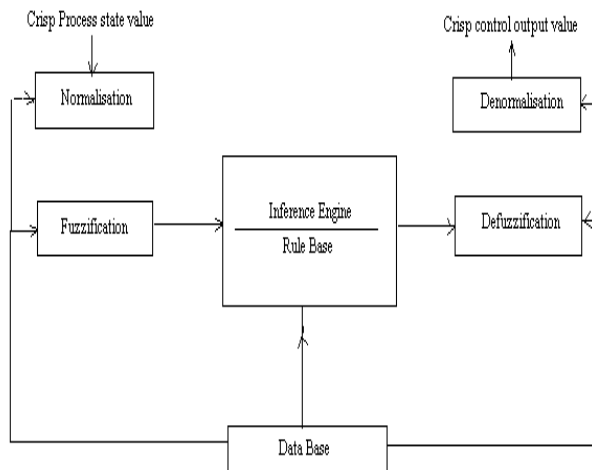


Fig 4 . Block Diagram of Fuzzy Logic Controller

Initial rule base that can be used in drive systems for a fuzzy logic controller consist of 49 linguistic rules, as shown in Table I, and gives the change of the output of fuzzy logic controller in terms of two inputs: the error (e) and change of error (de). The membership functions of these variables are given in Fig.5. In Table I, the following fuzzy sets are used: NB negative Big, NM

negative medium, NS negative small, ZR zero, PS positive small, PM positive medium and PB positive Big. For example, it follows from Table I that the first rule is:

IF **e** is NB and **de** is NB then **du** is NB

The linguistic rules are in the form of IF-THEN rules and take form:

IF (e is X and de is Y) then (du is Z),

Where X, Y, Z are fuzzy subsets for the universe of discourse of the error, change of error and change of the output. For example, X can denote the subset NEGATIVE BIG of the error etc. On every of these universes is placed seven triangular membership functions. It was chosen to set these universes to normalized type for all of inputs and output. The range of universe is set to -1 to 1.

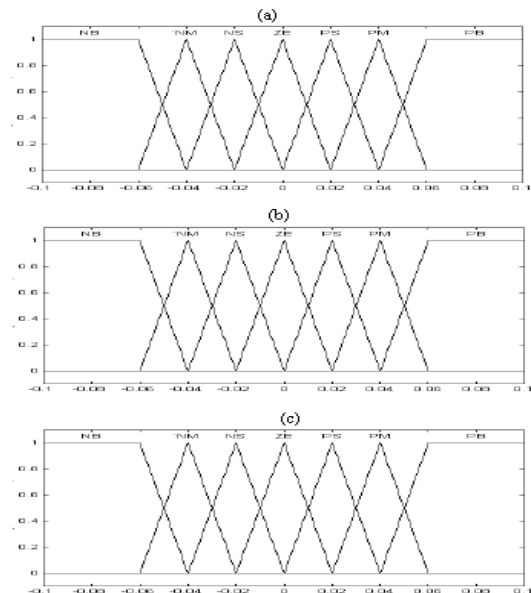


Fig 5. Membership function for error, error rate, controller output

D. THREE LEVEL INVERTER FED PMBLDC MOTOR WITH HYBRID CONTROLLER

The fuzzy PI controller is shown in Fig 6. The fuzzy controller is basically an input/output static nonlinear mapping, hence the controller action is in the form

$$K_1 E + K_2 CE = DU \quad (1)$$

Where K_1 and K_2 are nonlinear coefficients or gain factors.

$$DU = [K_1 E dt + K_2 CE dt] \quad (2)$$

$$U = K_1 [E dt + K_2 E] \quad (3)$$

Equation (3) is a fuzzy P-I controller with nonlinear gain factors.

TABLE I : FAM TABLE FOR HYBRID CONTROLLER

	NB	NM	NS	ZE	PS	PM	PB
NB	NB	NB	NB	NM	NS	NS	ZE
NM	NB	NM	NM	NM	NS	ZE	PS
NS	NB	NM	NS	NS	ZE	PS	PM
ZE	NB	NM	NS	ZE	PS	PM	PB
PS	NM	NS	ZE	PS	PS	PM	PB
PM	NS	ZE	PS	PM	PM	PM	PB
PB	ZE	PS	PS	PM	PB	PB	PB

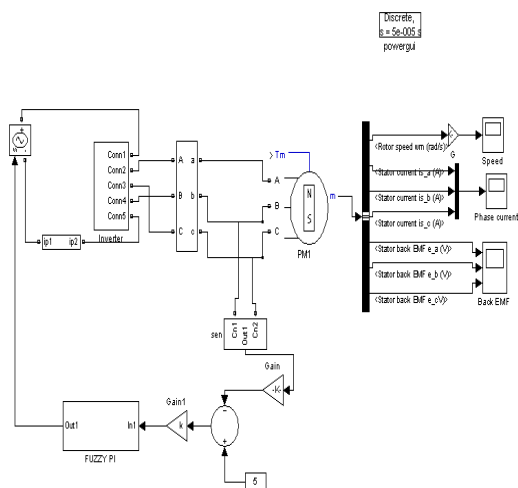


Fig. 6 Closed loop Sensorless Speed Control of PMBLDC motor using Hybrid Controller

III. SIMULATION RESULTS

A. RESPONSE OF HYBRID CONTROLLER

With the help of designed circuit parameters the MATLAB simulation of the above circuit is performed and the results are presented here.

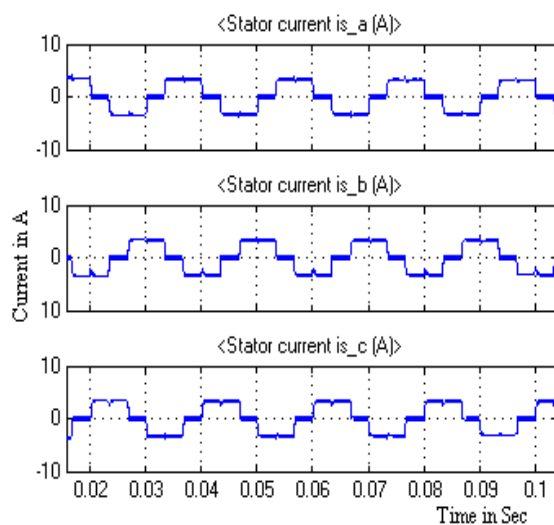


Fig 7. Three Phase Inverter Stator Current

The simulation results of stator current are shown in Fig 7. The currents are quasi square wave with a displacement of 120° .

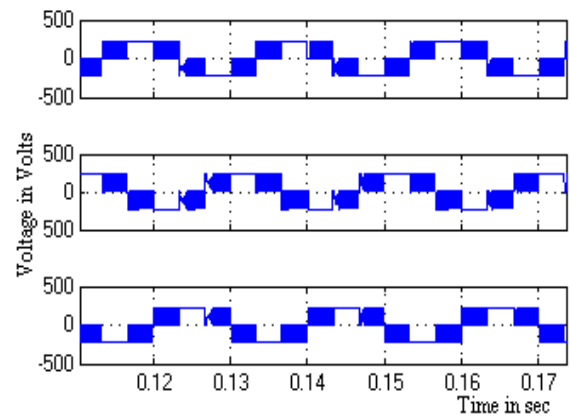


Fig 8. Three level Inverter Output Voltage

The stator voltages are shown in Fig 8. They are also displaced by 120° .

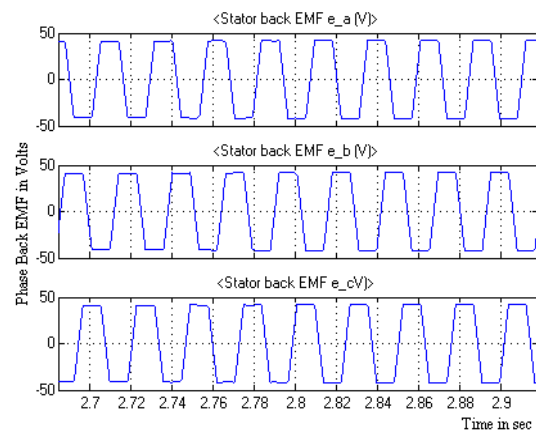


Fig 9. Trapezoidal shape Phase Back EMF

The stator phase back emf is shown in Fig 9. The phasor back emf is trapezoidal as shown.

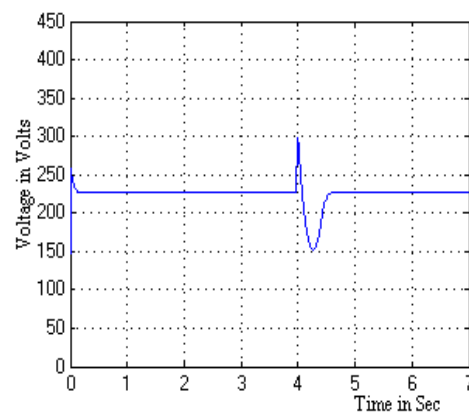
Fig.10 DC Input Voltage to Inverter with Disturbance at $t=4$ sec

Fig 10 shows the DC input voltage to the 3-level inverter with a disturbance at $t=4$ Sec. The closed loop system brings the voltage to the normal value by adjusting the input voltage of the inverter.

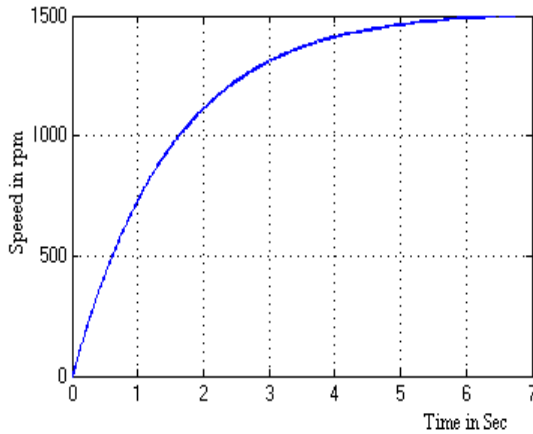


Fig. 11 Rotor Speed in rpm

The rotor speed characteristic of 3-level inverter fed PMBLDC motor using Hybrid controller is shown in Fig 11. The rotor reaches the rated speed in 5.5 Sec.

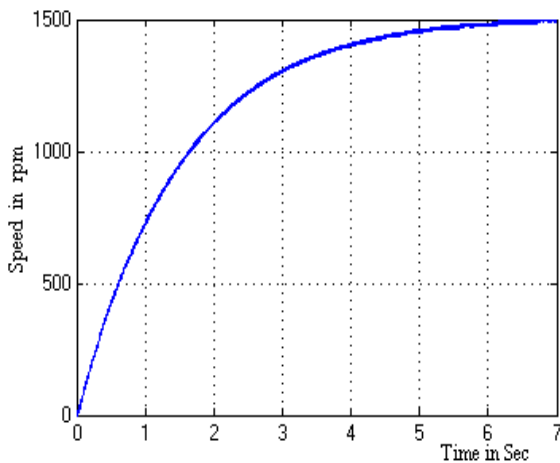


Fig 12. Rotor Speed in rpm

The rotor speed characteristic of 3-level inverter fed PMBLDC motor with PI controller is shown in Fig 12. The rotor reaches rated speed in 6.1 Sec. The time taken to settle at rated speed is comparatively more in conventional controller.

TABLE II PERFORMANCE ANALYSIS

Controller	THD	IAE	ISE	tss
PI Controller	0.5747	42.79	334	6.1
Fuzzy Controller	0.4173	28.39	107	6.00
Hybrid Controller	0.3825	10.49	29.35	5.5

IV. EXPERIMENTAL RESULTS

After the simulation studies, a Hybrid Fuzzy logic based Three level inverter fed PMBLDC motor is fabricated and tested. The top view of the hardware is depicted in Fig 13. The hardware consists of power circuit, control circuit and PMBLDC motor.

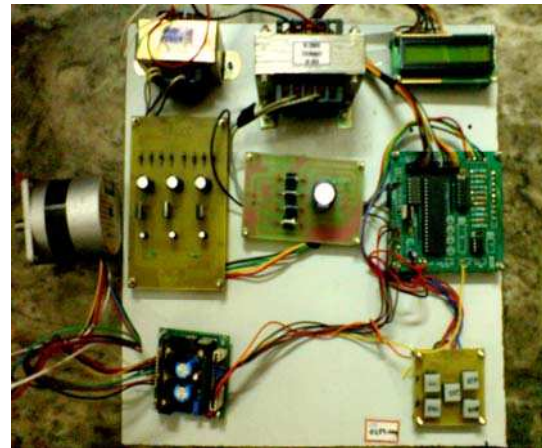


Fig 13 Top View of Hardware circuit

The regulators 7812 and 7805 in the control circuit give the DC supply required by the driver and microcontroller chips respectively. The driver chip amplifies 5V pulse to 10V level. DC output from the rectifier is ripple free due to the filter. The Atmel microcontroller 89C2051 is used to generate the pulses. Port 1 of the microcontroller is used for generating the gate pulses. Timer 0 is used for producing the delay required for the duration T_{ON} and T_{OFF} . The microcontroller operates at a clock frequency of 12 MHz.

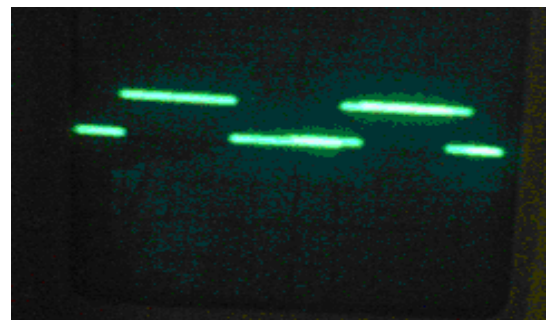


Fig 14 Pulse Signal waveform

The pulses produced by the microcontroller are amplified using the driver IC IR 2110. Three driver ICs are used to amplify the gate pulses. The oscillogram of pulse signal is given in Fig 14.

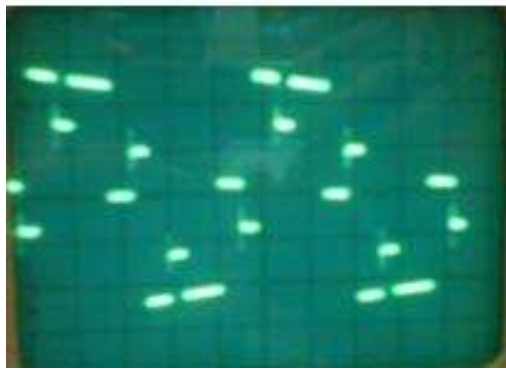


Fig 15 Three level inverter voltage

The oscillogram of three phase inverter output voltage is depicted in Fig 15. The Back EMF waveform of Three Phase PBLDC is depicted in Fig 16. The Back emf waveforms are trapezoidal as shown.

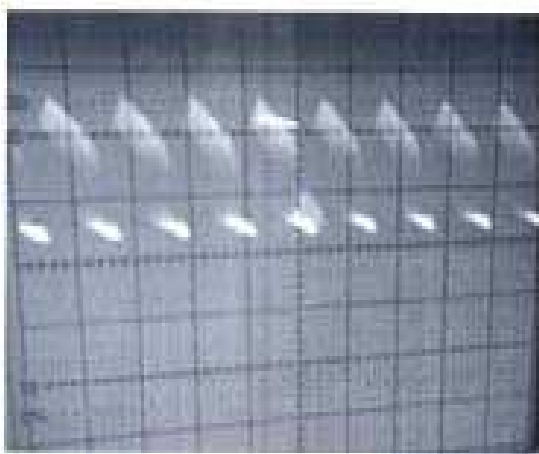


Fig16. Back EMF waveform

PMBLDC motors drives are used in a wide range of commercial and residential applications due to their highest possible efficiencies. The speed control ability of compressors and blowers is able to provide operation at their high efficiency. The physical integration of controller in the motor body itself is able to make them most suitable for low power (0.5hp) blowers and low power (50W) tube axial fans for cooling the electronic equipment.

V. CONCLUSION

The closed loop controlled sensorless PMBLDC drive is modeled and simulated using the blocks of simulink. The simulation results of closed loop system are presented. The closed loop system is able to maintain constant speed by maintaining constant voltage. The simulation results agree with the analytical predictions. The PMBLDCM drive system is successfully fabricated

and tested. The hardware system used in the present work has obvious advantage of using single phase supply. This drive can be used for variable speed applications like Electrical vehicles, Robotics etc., The experimental results coincide with the simulation results.

This Paper also presents a comparative study of fuzzy controllers with conventional controller of sensorless speed control of Permanent magnet Brushless DC Motor. The simulation results show that the Hybrid controller is the best performance in all aspects. It can be noticed that the hybrid controller exhibits fast rise time, no overshoot, and lesser settling time with lesser THD, IAE and ISE.

REFERENCE

- [1] B.K. Bose, Power Electronics and AC Drives, Prentice Hall, Englewood Cliffs, NJ: 1986.
- [2] L. Hao, H. A. Toliyat, "BLDC motor full-speed operation using hybrid sliding mode observer," in *Proc. IEEE-APEC Annu. Meeting*, Miami, FL, Feb. 9-13, 2003, vol. 1, pp. 286-293.
- [3] P. Pillay and R. Krishnan, "Application characteristics of permanent magnet synchronous and brushless dc motors for servo drives," *IEEE Trans. on Ind. Appl.*, vol. 27, no. 5, pp. 986-996, Sep./Oct. 1991.
- [4] T.J.E. Miller, "Brushless Permanent-Magnet and Reluctance Motor Drives," Oxford, 1989.
- [5] R. Krishnan, "Electric motor drives- modeling, analysis and control", Prentice Hall of India Private Limited, 2002.
- [6] K. Uzuka, H. Uzuhashi, et al., "Microcomputer Control for Sensorless Brushless Motor," *IEEE Trans. on Industry Application*, Vol. IA-21, May-June, 1985.
- [7] J. Shao, D. Nolan, and T. Hopkins, "A Novel Direct Back EMF Detection for Sensorless Brushless DC (BLDC) Motor Drives," *Applied Power Electronic Conference (APEC 2002)*, pp. 33-38.
- [8] N. Mastui, "Sensorless PM Brushless DC Motor Drives," *IEEE Trans. on Industrial Electronics*, Vol. 43, April 1996.
- [9] R. Krishnan, "Electric motor drives- modeling, analysis and control", Prentice Hall of India Private Limited, 2002.
- [10] Rene Zwahlen, Timothy Chang, "Feed forward speed control of Brushless DC motor with input shaping," The 33rd Annual Conference of the IEEE Industrial Electronics Society (IECON), Nov 5-8, 2007.
- [11] Mehdi Nasri, Hossein Nezamabadi-Pour, Malihemaghfoori, "A PSO-Based optimization of PID controller for a Linear BLDC Motor" *Proc. Of World academy of Science Engg & Tech*, Vol. 20, April 2007.
- [12] Zhiqiang Li & Changliangxia, "Speed Control of BLDC based on CMAC & PID controller" *Proc. Of 6th World congress on Intelligent Control & Automation*, China, June 21-23, 2006.
- [13] Yoko Amano, Toshio Tsuji, Atsushi Takahashi, Shigaonchi, "A Sensorless Drive system for BLDC using a Digital phase-Locked Loop," *Wiley periodicals Inc.* vol. 142 No. 1 pp. 1155-1162, 2003.
- [14] L.A. Zadeh, "Fuzzy sets," in *Information and Control*. New York: Academic, 1965, vol. 8, pp. 338-353.
- [15] "Outline of a new approach to the analysis of complex systems and decision processes," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, Jan. 1973.

- [16].C.Lee," Fuzzy logic in control systems, fuzzy logic controller, Parts I and II," IEEE Trans. Syst.,Man,Cybern., vol.20,pp.404-435,1990.
- [17].C.Elmas and O.F.Bay, " Modeling and operation of a nonlinear switched reluctance motor drive based on fuzzy logic," in Proc. Eur. Power Electronics Applications Conf., Sevilla, Spain, Sep.18-21, 1995, pp.3.592-3.597.
- [18].O.F.Bay, C.Elmas, and M.Alci," Fuzzy Logic based control of a switched reluctance drive," in Int. Aegean Conf. Electrical Machines Power Electronics, Kusadasi, Turkey, June 5-7, 1995, pp.333-337.
- [19].O.F.Bay, "Fuzzy control of a field orientation controlled induction motor," J.Polytechnic, vol.2, no.2, pp.1-9, 1999.
- [20].C.Elmas and M.A.Akcayol, "Fuzzy logic controller based speed control of brushless DC motor," J.Polytechnic, vol.3, no.3, pp.7-14, 2000.

T.V.Narmadha has obtained her AMIE (EEE) from the Institution of Engineers (India) and M.E., degree from College of Engineering, Anna University in the years 1996 and 2000 respectively. She has 11 years of teaching experience. She is presently a research scholar in Anna University. Her area of interest is Permanent Magnet Motor Drives.

T.Thyagarajan is the Professor and Head of the Department of Instrumentation Engineering, Anna University, Chennai, India. He obtained B.Tech degree from Government Engineering College, Anantapur; M.E degree from M.S. University, Baroda and Ph.D from Anna University. He also pursued Post-Doctoral Research at National Taiwan University, Taipei. He has received DTE Award for guiding best U.G. project and SISIR Kumar award for publishing best research paper. He made technical visits to many countries. He is a member of IEEE, IE(I), ISTE, SSI and ISO(I). His teaching and research interests include electrical drives, auto tuning, process modeling and control using fuzzy logic, neural networks and genetic algorithm, sensor networking and soft sensors.

On Performance of Multicast Delivery with Fixed WiMAX Telemedicine Networks Using Single-Carrier Modulation

Bernard Fong

Prognostics and Health Management Center, City University of Hong Kong
Email: bfong@ieee.org

Guan Yue Hong

Email: gyh138@gmail.com

Abstract—IEEE 802.16e fixed WiMAX provides a low cost solution for integrated multimedia access networks with a wide bandwidth making it particularly suitable for telemedicine applications. While a wide variety of modulation schemes are suitable for fixed wireless systems, single carrier modulations such as QAM offer advantages such as reduction in power requirements and system complexity compared to multi-carrier transmission systems. In this paper, we analyze the performance of QAM schemes used in a fixed WiMAX system that supports multicast distribution of real-time traffic for healthcare services by evaluation of system performance on a 10 GHz carrier. Results are presented by comparing the distribution of video data using QPSK and 16-QAM and bandwidth utilization is calculated for continuous data transmission in remote patient monitoring.

Index Terms— broadband access, modulation, QAM, telemedicine, WiMAX

I. INTRODUCTION

IEEE 802.16e WiMAX networks are widely used for providing point-to-multipoint network access for fixed locations within a radius of several kilometers due to its design for low cost two-way transmission. Its main advantages include ease of expansion and allowing frequency reuse. Fixed WiMAX provides a means of wireless data distribution at high data rates over a reasonably large area compared to alternative solutions such as wireless local loop (WLL) and wireless cable (multichannel multipoint distribution system), transmission of various types of medical data at frequencies in excess of 10 GHz requires high resolution planning since data packets are vulnerable to burst errors.

Wireless communication systems are affected by channel-induced phenomena such as fading that can

significantly impact their performance. In wireless communication systems, a base station modem unit (BMU) is connected to each base station controller (BSC). Its purpose is to convert the digital data to be carried across the radio interface into a format appropriate for transmission over wireless channels. Different equipment manufacturers employ various modulation schemes for their BMU implementations as they have different standards. While the primary concern is making a decision on compromise between cell coverage and data throughput, there is no straightforward answer as to which modulation scheme is best for a wireless network. The optimal tradeoff depends mainly on specific application, for example, systems for life-saving missions would have far more stringent requirements than those for general health assessment. In this paper, we investigate the suitability of using a wireless point-to-multipoint system for distribution of multimedia traffic based on a QAM scheme that is optimized for providing telemedicine services within a local environment [1].

QAM is a very robust type of modulation scheme that provides a high capacity. However, QAM systems are subject to fast Rayleigh fading and time delay spread. Compensation for distortion over Rayleigh fading channels has been studied [2]. In [3], a discussion about the effects of modulated signals transmitted over a fading channel is presented which also includes an outline of a numerical evaluation on such effect [4]. While QAM offers a comparably high number of bits per hertz [5], various QAM modulation schemes have been used by different equipment manufacturers according to some tradeoff between bits per Hertz and bits per unit area coverage. High modulation schemes such as a 128-QAM offers more bits per hertz at the expense of cell size reduction. Numerous research results [6], [7], [8] have been reported with a number of receivers studied illustrating various receiver structures. The receiver structures required for high modulation QAM schemes

Manuscript received October 1, 2009; revised December 23, 2009; accepted January 21, 2010.

Corresponding author: Bernard Fong, Prognostics and Health Management Center, City University of Hong Kong, 83 Tat Chee Road, Kowloon, Hong Kong, Tel: +852 3442 5936 Email: bfong@ieee.org

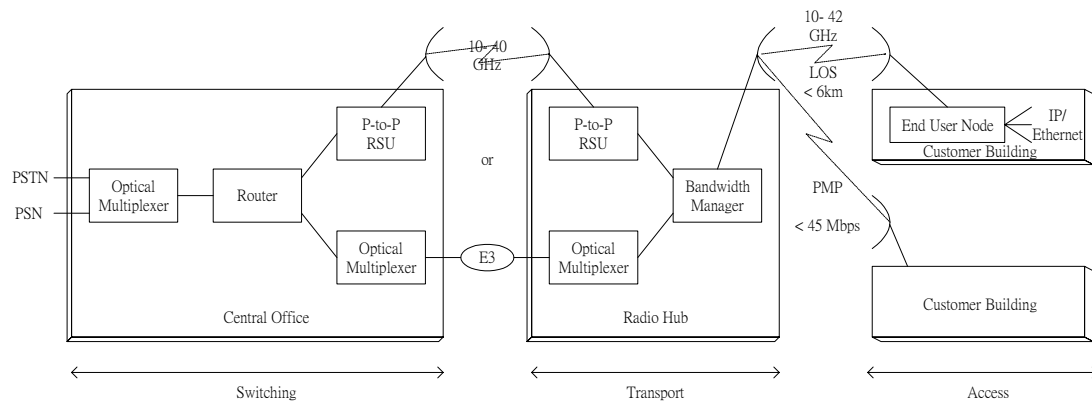


Figure 1. A point-to-multipoint (PMP) network infrastructure

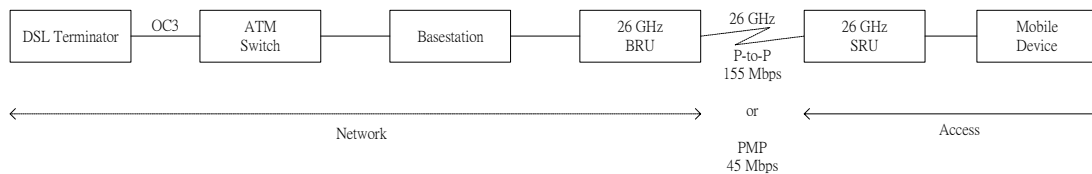


Figure 2. System block diagram

become much more complex than those required for 4 (QPSK) and 16-QAM schemes. Further, an adaptive equalization scheme for a 16-QAM receiver has been proposed [9]. Low modulation QAM schemes are therefore very attractive in terms of costs and size of receiving devices used by consumers. To minimize the effects of cell-to-cell interference, lower modulation schemes are studied.

The use of Orthogonal Frequency Division Multiplexing (OFDM) techniques for multiplexed QAM for a voiceband modem had been studied extensively in the 1980's [10], [11]. More recently, the authors of [12] employed a Coded Orthogonal Frequency Division Multiplexing (COFDM) transmission scheme for an IEEE 802.16 based local multipoint distribution systems (LMDS), leading to further research opportunities for development of modems using an OFDM transmission scheme. The transmission channel is characterized in terms of time and frequency fading predominantly due to movement of receivers relative to the transmitter and multipath propagation, respectively. Although OFDM exhibits certain advantages, single carriers such as variants of QAM offer comparable advantages such as lower power efficient which is particularly suitable for wearable health monitoring receivers; asymmetrical operations is made easier due to simpler transmission circuitry relative to that of reception. Further, high level of narrow-band noise immunity due to inherent capability by use of adaptive equalization makes QAM particularly suitable for transmission over such systems.

We evaluate the system performance by simulating its bit-error probability and coverage. In our experiments, we assumed that the video transmitted does

not contain redundancies such as error correction or synchronization characters. It is further assumed that the data traffic volume associated with such redundancies is very much less than that of the video data itself hence the omission would not affect actual system performance.

In this introductory section, we have summarized the advantages of using single-carrier modulation techniques such as QAM and stated the differences between using a high (e.g. $M=128$) and a low (e.g. QPSK, $M=4$) order M -ary QAM scheme. The remaining sections of this paper are organized as follows. In Section II, we describe the system layout for performance evaluation and the OFDM transmission technique over fading channels is discussed with its performance analyzed in Section III. In Section IV, channel utilization is discussed with analysis on sending multimedia data over the same channel simultaneously. Finally, we conclude the paper in Section V.

II. SYSTEM LAYOUT

A. Set Up

Position A network has been installed to provide broadband wireless access (BWA) for data delivery through wireless networks across premises of close proximity. While different countries have different spectrum allocations with appropriate regulations, they generally operate in the range of 10 to 66 GHz, and mostly below 40 GHz. In places where radio link availability is greatly affected by persistent heavy rainfall, such as those classified as region-P by ITU, a lower frequency of around 10 GHz is preferred as it is

less affected by rain attenuation. Fig. 1 shows the basic operation of a typical network currently set up around the world to provide broadband wireless access delivering a range of services such as video multicasting [13].

The system consists of three main components. It uses the switching and transport portions of the network for connection to the backbone network that processes the incoming multimedia data for distribution over the wireless channel. The system consists of a DSL terminator equipped with an OC3-ATM card used to route IP traffic over ATM connected to the network backbone. The base station that controls the base station radio unit (BRU) consists of a base modem module, a base network module, and a base station controller. The BRU is a self-contained unit with a 45-degree antenna azimuth at 7-degree elevation. The access side, separated by an outdoor wireless channel from the system, evaluates the system performance by studying the received signal. In this network, video data is distributed to make use of wireless internet access which transmits both the video and audio (sound track) signals over wireless broadband channels. Our system is initially intended for stationary reception due to bulkiness of the receiver. However, further research on optimizing for mobile receivers traveling at high speed will be conducted for extending services to ambulances.

Fig. 2 shows a schematic diagram of its operations while keeping the basic elements of the network infrastructure unchanged. The system is designed to provide a range of multimedia services for subscribers through wireless channels. It is optimized for reliable transmission of video with other forms of data such as text and graphics handled at a much lower priority.

The video data is played and distributed to the transmitter through the switching center and up to the wireless backbone network. The BMU handles data traffic between subscribers and the network. In our system, it serves as a modulator that processes the video clip data and sends it via the radio channel to the subscriber receivers. In this system, the BMU does not receive anything from the subscribers as we made an assumption that the data is sent in simplex mode. The envelope of the received signal is given by

$$r(t) = a(t) \cdot \sum_{k=1}^N s_k(t - kT) \quad (1)$$

Where $s(t)$ is a signaling pulse and T is the symbol period; N is the total number of paths and $k=1$ is the line-of-sight (LOS) path. The distortion caused by multipath fading $a(t)$ with linearity [14], [15] is given by:

$$a(t) = a_k(0) + a_k(1) \frac{t - kT}{T} \quad (2)$$

B. Modulation Scheme

Factors that influence our choice of modulations scheme are:

1. Fading immunity
2. Spectral efficiency
3. Receiver complexity
4. Power efficiency

As discussed in Section I, single-carrier modulation offers better overall compromise for use in many wireless systems. Our work therefore concentrates on comparing lower order QAM schemes such as QPSK and 16-QAM.

System immunity to fading is important to ensure reliable operation under both time and frequency selective fading environments. The effects of fading are discussed in Section II D. Spectral efficiency is a measure of data rate per unit bandwidth per unit area given by bps/Hz/m². Our aim is to maximize this quantity. Receiver complexity and power efficiency are important considerations, particularly for mobile receivers. For example, power efficiency places a limitation on battery size for mobile receivers.

In essence, higher order M-ary QAM schemes (e.g. M=128) provide higher spectral efficiency at the expense of receiver sensitivity with an increase in modulation level and poor noise immunity. A tradeoff between bps/Hz and transmission quality is therefore an issue in comparison of QAM schemes [16]. We present the result from our study of this tradeoff in Section III.

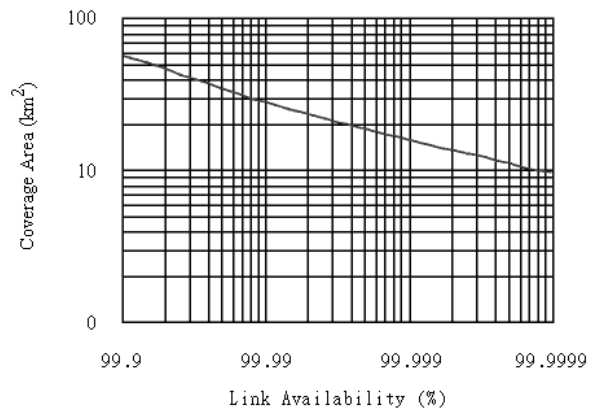


Figure 3. Coverage vs. link availability

C. Link Availability

The effect of path loss reduces signal power over distance. The maximum range that indicates the maximum distance of receivers from the basestation, depends primarily on antenna gain and rainfall statistics. The fade margin can be adjusted for higher link availability. The range covered by a 10 GHz carrier as a function of link availability in percentage of time is illustrated in Fig. 3. The maximum range decreases approximately linearly with a higher availability. The link with 99.99% availability, disabled for no more than

52 minutes per year, offers a range of close to 20km. This is to ensure a bit error rate (BER) performance of 10^{-6} or below. Result shown indicates the maximum link range for a transmission rate of 12 Mbps with 10 GHz carrier with LOS and no rainfall. The range for 99.99% availability is 18 km.

D. Effects of Generalized Fading Channel

In a fixed network environment, the main sources of signal distortion are

- 1) Multipath fading
- 2) Frequency selective fading
- 3) Additive noise

Multipath fading is a collective term used to describe the constructive and destructive superposition of signal components that have taken different paths due to such phenomena as scattering, diffraction and reflection. It therefore amounts to time selective fading. The signal propagation path often has no direct LOS due to physical obstacles between transmitter/receiver antennas. The spectral components of the signal due to frequency selective fading are affected by different fading amplitudes and phase shifts as the signal bandwidth is larger than that of the channel's coherence bandwidth. We consider the effects of generalized channel fading below.

Although in conditions where severe interference and fading make robust carrier recovery become a necessity, the use of OFDM offers good performance in fading and time-variant transmission media [17]. Experiment shows that low-order QAM schemes also provide adequate performance in such situation. A comparison is carried out between the ideal channel and our estimated channel based on assumptions made in [18], [19]. The receiver remains stationary and the bit rate is varied with a carrier frequency of 10 GHz. The channel is defined with the properties

$$pdf(R) = \frac{2m^m R^{2m-1}}{R^{2m}} \cdot e^{-m} \quad (3)$$

where R is a random variable representing the mean-square fading amplitude and

$$m = \frac{R^2}{\text{var}(R^2)} \quad |m| \geq 0.5$$

There is no fading when m tends to infinity as the pdf becomes an impulse function. The received signal is subjected to additive white Gaussian noise (AWGN), which is assumed to be independent of channel fading. Fig. 4 shows the performance of the receiver with $M = 4$ and $M = 16$.

We consider a multilink channel having L independent fading channels as described in [20], the equivalent low-pass impulse response of the time

dispersion as the signal propagates through the frequency selective fading channel is given by:

$$h(t) = \sum_{l=1}^L a_l \cdot e^{-j\theta} \cdot \delta(t - \tau) \quad (4)$$

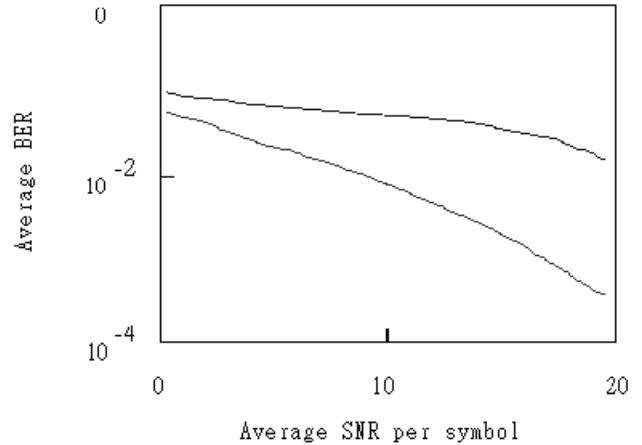


Figure 4. Average BER performance of receiver

where $\delta(t-\tau)$ is the Dirac delta function, L is the number of paths with the first LOS path being $L = 0$ and a , θ , and τ represent the amplitude, phase, and magnitude of time delay for each path, respectively.

III. PERFORMANCE ANALYSIS

The purpose of our experiments is to evaluate the E_b/N_0 performance of QPSK and 16-QAM for our system under the transmission channel with characteristics described earlier in II D. To maintain link availability, the comparison is made at the cut-off point at which the link is no longer available where BER reaches 10^{-6} . Gray encoding with absolute phase coherent detection has been used to improve BER performance. From Fig. 5, QPSK shows a better E_b/N_0 performance of 4.3 dB over 16-QAM. The symbol-error rate (SER) of each symbol is different between $M = 4$ and $M = 16$. The measured system data rate is 12 Mbps.

Thus, 16-QAM performs noticeably poorer compared to QPSK as shown in Fig. 5. Larger M may offer better performance in number of bits per baud at the expense of increased receiver structure complexity and decrease in BER performance. The system with $M = 4$ and $M = 16$ has been compared under different symbol lengths and its performance is shown in Fig. 6. It shows that both 4 and 16-QAM perform very similarly with a constant E_b/N_0 value and varying symbol rate.

Results for BER better than 10^{-6} are not presented as the link is considered available and analysis is assumed to be the worst scenario. In actual fact, the difference in performance widens between these modulation schemes as BER decreases further.

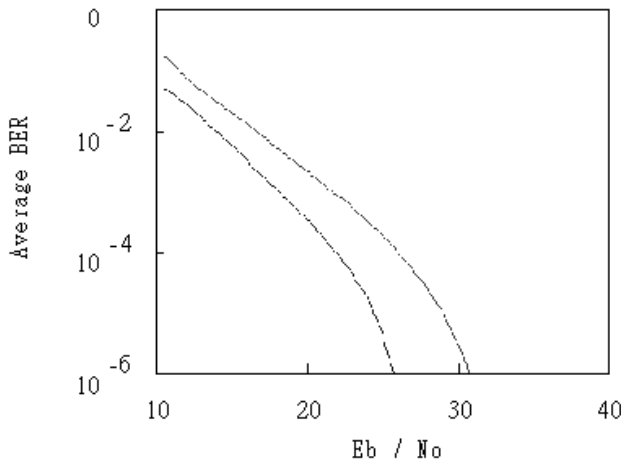


Figure 5. BER comparison between QPSK and 16-QAM

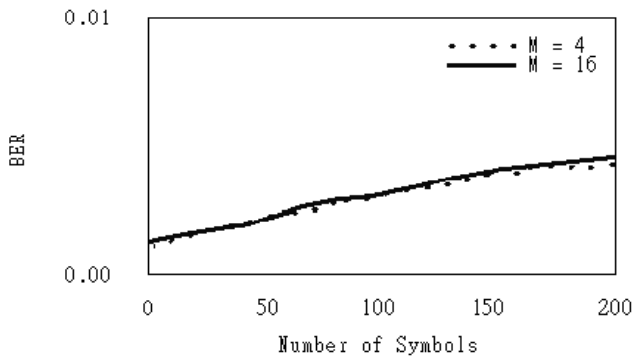


Figure 6. BER vs. data symbol length

Analysis shows that increasing the modulation order from 4 to 16 offers a very slight improvement in bit error rate. However, QPSK offers a slight advantage of simpler receiver structure for the customer modem at the expense of decrease in spectral efficiency by a ratio of 15:3.5. Although it appears that QPSK offers a better compromise than 16-QAM when data symbol length is varied.

IV. BANDWIDTH UTILIZATION

The system has been set up to measure the suitability of various modulation techniques. With data collected from the actual system, a series of computer simulations are carried out to examine the system's behavior under different contents of audio and video data transmitted. The importance of bandwidth utilization in such network impacts subsequent stages of data processing. Electronic Patient Record (EPR) updating can be affected when the network saturates. The consequential network degradation resulting from saturation can be mitigated by automatic fuzzy ontology algorithms similar to that in [21].

To evaluate the effects of network saturation, we simulate the channel of 9.9 – 10.5 GHz when shared by 5

subscribers, n is defined as the number of subscribers sharing the channel (i.e. $n = 5$ in our simulation). Each subscriber uses the channel to send different types of data to the BMU. The channel bandwidth is dynamically assigned to utilize the allocated bandwidth to each subscriber. With a mixture of audio (voice) and video (series of images at a fixed rate of 29.7 frames per second) being used as test data, the redundancy bandwidth for audio signal is assumed to be 60% of the time as proposed by [22]. A protocol described in [23] uses redundancy when there is no voice signal transmitted, as detected by voice activation, to transmit other types of signals. With a channel capacity of 50 cells per block for data with overheads neglected, we compute the system throughput as a variation of redundancy bandwidth.

Cell delay is the combined effect of BMU cell waiting time and the actual data transmission time, and the system throughput is measured by the average number of cells transmitted per block of the channel. We assume that video transmission is constant at 29.7 fps with no redundancy, and we further assume that the only data sent is video composed of a mono audio track and pictures only. The system throughput S is given by

$$S = S_A + S_V \quad (5)$$

An assumption that the system only handles data traffic consists of audio and video without other information such as additional synchronization or error correction redundancies has been made so that data handled by the system consists only of audio and video without any other types of data.

For the characteristics of video, its throughput S_V is assumed to be constant which is determined by a function of the transmission rate R_V is given by

$$R_V = P_h \cdot P_v \cdot c \cdot f \quad (6)$$

where P_h and P_v are the number of horizontal and vertical pixels respectively, c is the color depth (e.g. 8 bits for 2^8 or 256 colors) and f is the scan rate in number of frames per second (typically ~30fps for video clips). In our simulation, we tested a video sample of resolution 320 x 240 with 8-bit or 256 colors at 29.7 fps. In this case, R_V is 4.8 Mb/s.

The throughput characteristics of audio S_A is given by

$$S_A = R_A \cdot n \cdot K \cdot \frac{T_t}{T} \quad (7)$$

$$\text{where } T = T_t + T_i$$

The following parameters are used to compute the throughput as an average number of cells transmitted per block of the channel. R_A is the data rate or the bandwidth assigned for the audio (voice) signal. K is the simulation parameter for the audio signal that measures the

proportion of time that there is audio data to be sent ($0 \leq K \leq 10$). T is the total transmission time that derives the redundancy bandwidth as in [24] such that T_i and T_i denote the duration of sound being transmitted and the duration when there is no sound (idle), respectively.

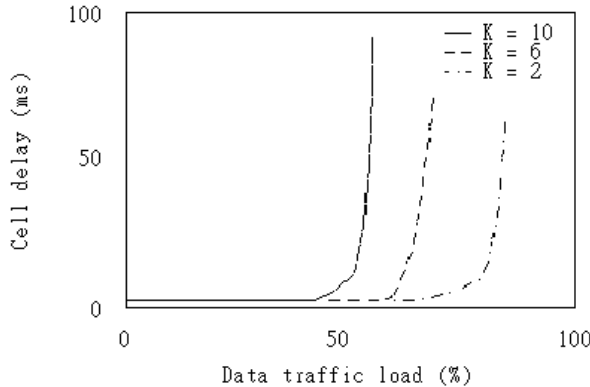


Figure 7. Cell delay characteristics of transmitted data

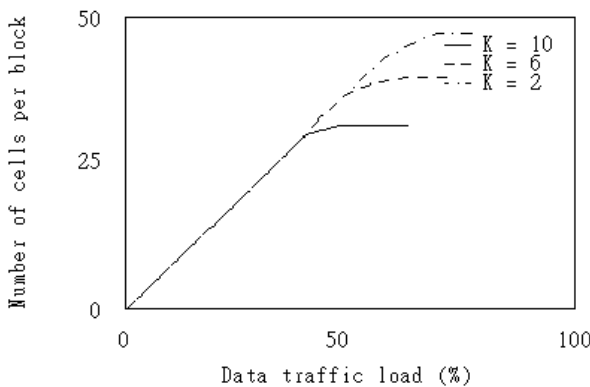


Figure 8. System throughput characteristics

Fig. 7 shows that each cell maintains a low latency and a small cell delay. When more audio data is transmitted (i.e. K increases and R_A also increases), cell delay becomes more severe. Fig. 8 shows the throughput characteristics of the system where its throughput increases with data traffic load linearly until its capacity approaches the maximum available bandwidth. It is noted from Fig. 8 that the channel utilization is affected by T_i as a large T_i decreases channel utilization. So, R_A determines the bandwidth efficiency where it is a close approximation to the system maximum throughput.

V. CONCLUSIONS

We have compared two single-carrier modulation schemes for good compromise between bandwidth efficiency and ease of implementation for use with a fixed WiMAX telemedicine system at a carrier frequency of 10 GHz for distribution of multimedia data over a local area. Although lower modulation schemes offer a reduction in receiver structure complexity at the expense

of significant degradation in E_b/N_0 performance, this paper leads to the conclusion that a 16-QAM scheme provides optimal compromise between data throughput and cell coverage while provides adequate performance for multicast distribution of video traffic. It is also apparent that the objective of an efficient deployment would be to maximize bandwidth utilization by considering the proportion of time that the audio track is silent.

REFERENCES

- [1] B. Fong and G. Y. Hong, "RF net scales broadband to local area", EETimes, June 17, 2002.
- [2] J. K. Carvers, "An analysis of pilot symbol assisted modulation for Rayleigh Fading Channels", Proc. IEEE VTC, St. Louis, pp. 380-385, 1991
- [3] M. S. Alouini and Goldsmith A. J., "A Unified Approach for Calculating Error Rates of Linearly Modulated Signals over Generalized Fading Channels", IEEE Trans. Communications, Vol. 47 No. 9, pp. 1324-1330, Sep 1999
- [4] A. C. M. Fong, S. C. Hui and C. T. Lau, "An experimental learn-on-demand system in a wireless campus environment", IEEE Multimedia, Vol. 11/4, pp. 50-60, 2004.
- [5] I. Korn, "Digital Communications", Van Nostrand Reinhold, 1985
- [6] V. Mignone, and A. Morello, "CD3-OFDM: a novel demodulation scheme for fixed and mobile receivers", IEEE Trans. Communications, Vol. 44 No. 9, pp. 1144- 1151, Sep 1996
- [7] B. D. Hart and D. P. Taylor, "Extended MLSE receiver for the frequency-flat, fast-fading channel", IEEE Trans. Communications, Vol. 46 No. 2, pp. 381- 389, May 1997
- [8] A. Mouaki Benani and F., Gagnon, "Comparison of carrier recovery techniques in M-QAM digital communication systems", Proc. Electrical and Computer Engineering Conf., Canada, Vol. 1, pp. 73- 77, 7- 10 Mar 2000
- [9] R. A. Peloso, "Adaptive equalization for advanced television", IEEE Trans. Consumer Electronics, Vol. 38 No. 3, pp. 119- 126, Aug 1992
- [10] B. Hirosaki, "An Orthogonally multiplexed QAM system using the discrete Fourier Transform", IEEE Trans. Communications, Vol. 29, pp. 982- 989, Jul 1981
- [11] S. Hirosaki et. al., "Advanced groupband data modem using orthogonally multiplexed QAM technique", IEEE Trans. Communications, Vol. 34, pp. 587-592, Jun 1986
- [12] V. Tralli et. al., "Adaptive Time and Frequency Resource Assignment with COFDM for LMDS Systems", IEEE Trans. Communications, Vol. 49 No. 2, pp. 235-238, Feb 2001
- [13] "Local Multipoint Distribution System", Web ProForum Tutorials of The International Engineering Consortium, <http://www.iec.org>
- [14] G. M. Vitetta, D. P. Taylor and U. Mengali, "Double-filtering receivers for PSK signals transmitted over Rayleigh frequency-flat fading channels", IEEE Trans. Communications, Vol. 44 No. 6, pp. 686-695, Jun 1996
- [15] G. M. Vitetta, U. Mengali and D. P. Taylor, "Double-filter differential detection of PSK signals transmitted over linearly time-selective Rayleigh fading channels", IEEE Trans. Communications, Vol. 47 No. 2, pp. 239-247, Feb 1999
- [16] S. Sampei, "Applications of Digital Wireless Technologies to Global Wireless Communications", Prentice-Hall NJ, 1997

- [17] W. Y. Zou and Y. Wu, "COFDM: An Overview", IEEE Trans. Broadcasting, Vol. 41 No. 1, pp. 1-8, Mar 1995
- [18] M. S. Alouini and A. J. Goldsmith "A Unified Approach for Calculating Error Rates of Linearly Modulated Signals over Generalized Fading Channels", IEEE Trans. Communications, Vol. 47 No. 9, pp. 1324-1334, Sep 1999
- [19] M.A. Do and S.Y. Wu, "Hybrid diversity combining techniques for DS-CDMA over a multipath fading channel", ACM Wireless Networks 3, pp. 155-158, 1997
- [20] F. Classen and H. Meyr, "Frequency synchronization algorithms for OFDM systems suitable for communication over frequency-selective channels", Proc. VTC, pp. 1655-1659, 1994
- [21] Q. T. Tho, S. C. Hui, A. C. M. Fong and T. H. Cao, "Automatic fuzzy ontology generation for semantic web", IEEE Trans Knowledge and Data Engineering, Vol. 18/6, pp. 842-856, June 2006.
- [22] P. T. Brady, "A Statistical Analysis of on-off patterns in 16 conversions", Bell Syst Tech. J., Vol 47, pp. 71-93, Jan 1968
- [23] K. S. Meier-Hellstern, G. P. Pollini, D. J. Goodman, "Network protocols for the cellular packet switch", IEEE Trans. Communications, Vol. 47, pp. 1235- 1244, Feb-Apr. 1994
- [24] P. H. Moose, "A technique for orthogonal frequency division multiplexing frequency offset correction", IEEE Trans. Communications, Vol. 42 No. 10, pp. 1590-1598, Oct 1994



Bernard Fong is currently with the Prognostics and Health Management Center, City University of Hong Kong. He received his BS degree from University of Manchester Institute of Science and Technology, United Kingdom, and PhD in healthcare information systems from the University of New South Wales, Australia. Prior to joining CityU PHM Centre (PHMC), he was a Visiting Associate Professor with the Hong Kong Polytechnic University. His current research interests are broadly in the areas of biomedical engineering, health informatics, and prognostics for healthcare and consumer electronics applications. Dr. Fong is serving as the South Asia Representative of the IEEE International Conference on Consumer Electronics (ICCE) and General Chair for the ICST Connecting Health International Workshop, co-sponsored by the ACM and IEEE-EMBS; he is currently an editorial board member of the Journal of Advances in Information Technology, IEEE Consumer Electronics Newsletter, and held a number of other positions such as Editor-in-Chief for the International Journal of Information Technology, Associate Editor for the Electronic Commerce Research & Applications Journal, and Guest Editor for special feature topic in wireless telemedicine of the IEEE Communications Magazine. He is authoring the book Telemedicine Technologies: Information Technologies in Medicine and Telehealth, to be published by John Wiley & Sons U. K. (2010).



Guan Yue Hong is an IT consultant.

Call for Papers and Special Issues

Aims and Scope

JAIT is intended to reflect new directions of research and report latest advances. It is a platform for rapid dissemination of high quality research / application / work-in-progress articles on IT solutions for managing challenges and problems within the highlighted scope. JAIT encourages a multidisciplinary approach towards solving problems by harnessing the power of IT in the following areas:

- **Healthcare and Biomedicine** - advances in healthcare and biomedicine e.g. for fighting impending dangerous diseases - using IT to model transmission patterns and effective management of patients' records; expert systems to help diagnosis, etc.
- **Environmental Management** - climate change management, environmental impacts of events such as rapid urbanization and mass migration, air and water pollution (e.g. flow patterns of water or airborne pollutants), deforestation (e.g. processing and management of satellite imagery), depletion of natural resources, exploration of resources (e.g. using geographic information system analysis).
- **Popularization of Ubiquitous Computing** - foraging for computing / communication resources on the move (e.g. vehicular technology), smart / 'aware' environments, security and privacy in these contexts; human-centric computing; possible legal and social implications.
- **Commercial, Industrial and Governmental Applications** - how to use knowledge discovery to help improve productivity, resource management, day-to-day operations, decision support, deployment of human expertise, etc. Best practices in e-commerce, e-commerce, e-government, IT in construction/large project management, IT in agriculture (to improve crop yields and supply chain management), IT in business administration and enterprise computing, etc. with potential for cross-fertilization.
- **Social and Demographic Changes** - provide IT solutions that can help policy makers plan and manage issues such as rapid urbanization, mass internal migration (from rural to urban environments), graying populations, etc.
- **IT in Education and Entertainment** - complete end-to-end IT solutions for students of different abilities to learn better; best practices in e-learning; personalized tutoring systems. IT solutions for storage, indexing, retrieval and distribution of multimedia data for the film and music industry; virtual / augmented reality for entertainment purposes; restoration and management of old film/music archives.
- **Law and Order** - using IT to coordinate different law enforcement agencies' efforts so as to give them an edge over criminals and terrorists; effective and secure sharing of intelligence across national and international agencies; using IT to combat corrupt practices and commercial crimes such as frauds, rogue/unauthorized trading activities and accounting irregularities; traffic flow management and crowd control.

The main focus of the journal is on technical aspects (e.g. data mining, parallel computing, artificial intelligence, image processing (e.g. satellite imagery), video sequence analysis (e.g. surveillance video), predictive models, etc.), although a small element of social implications/issues could be allowed to put the technical aspects into perspective. In particular, we encourage a multidisciplinary / convergent approach based on the following broadly based branches of computer science for the application areas highlighted above:

Special Issue Guidelines

Special issues feature specifically aimed and targeted topics of interest contributed by authors responding to a particular Call for Papers or by invitation, edited by guest editor(s). We encourage you to submit proposals for creating special issues in areas that are of interest to the Journal. Preference will be given to proposals that cover some unique aspect of the technology and ones that include subjects that are timely and useful to the readers of the Journal. A Special Issue is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

The following information should be included as part of the proposal:

- Proposed title for the Special Issue
- Description of the topic area to be focused upon and justification
- Review process for the selection and rejection of papers.
- Name, contact, position, affiliation, and biography of the Guest Editor(s)
- List of potential reviewers
- Potential authors to the issue
- Tentative time-table for the call for papers and reviews

If a proposal is accepted, the guest editor will be responsible for:

- Preparing the "Call for Papers" to be included on the Journal's Web site.
- Distribution of the Call for Papers broadly to various mailing lists and sites.
- Getting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Instructions for Authors.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

Special Issue for a Conference/Workshop

A special issue for a Conference/Workshop is usually released in association with the committee members of the Conference/Workshop like general chairs and/or program chairs who are appointed as the Guest Editors of the Special Issue. Special Issue for a Conference/Workshop is typically made of 10 to 15 papers, with each paper 8 to 12 pages of length.

Guest Editors are involved in the following steps in guest-editing a Special Issue based on a Conference/Workshop:

- Selecting a Title for the Special Issue, e.g. "Special Issue: Selected Best Papers of XYZ Conference".
- Sending us a formal "Letter of Intent" for the Special Issue.
- Creating a "Call for Papers" for the Special Issue, posting it on the conference web site, and publicizing it to the conference attendees. Information about the Journal and Academy Publisher can be included in the Call for Papers.
- Establishing criteria for paper selection/rejections. The papers can be nominated based on multiple criteria, e.g. rank in review process plus the evaluation from the Session Chairs and the feedback from the Conference attendees.
- Selecting and inviting submissions, arranging review process, making decisions, and carrying out all correspondence with the authors. Authors should be informed the Author Instructions. Usually, the Proceedings manuscripts should be expanded and enhanced.
- Providing us the completed and approved final versions of the papers formatted in the Journal's style, together with all authors' contact information.
- Writing a one- or two-page introductory editorial to be published in the Special Issue.

More information is available on the web site at <http://www.academypublisher.com/jait/>.