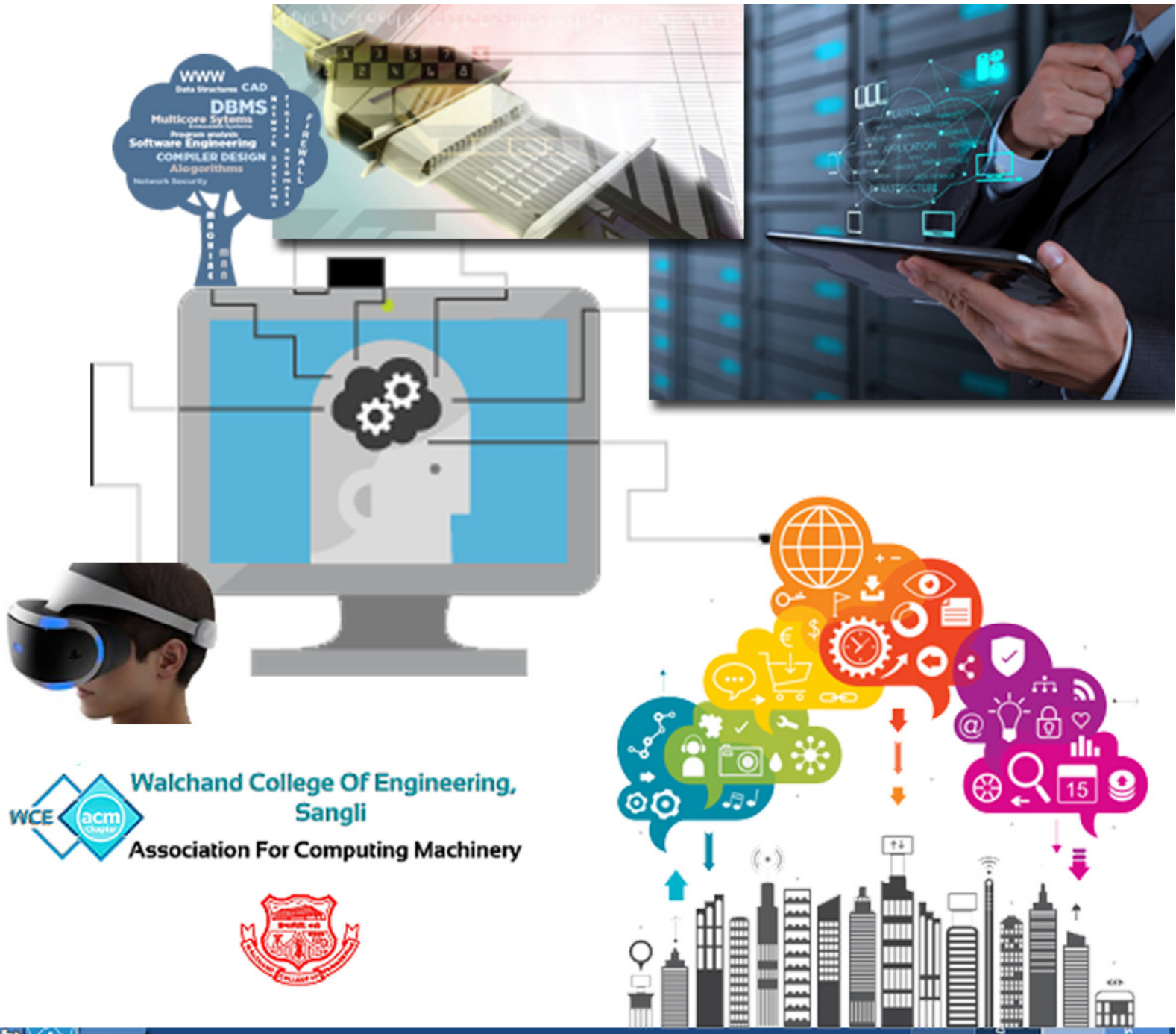
e-Proceeding of

National Research Symposium on Computing - RSC 2016

Jointly organized by
WCE ACM Student Chapter,
Department of Computer Science & Engineering and Department of Information Technology
Walchand College of Engineering, Sangli.



Walchand College Of Engineering, Sangli
Association For Computing Machinery

WALCHAND COLLEGE OF ENGINEERING

WCE

Content

Key-Note Speeches

Oral Presentation Papers

Poster Presentation Papers

ORGANIZING COMMITTEE

Sr. No	Name	Designation
1	Dr. P. J. Kulkarni	*Deputy Director & Professor - Department of Computer Science & Engineering, Walchand College of Engineering, Sangli*
2	Dr. B. F. Momin	*Associate Professor - Department of Computer Science & Engineering, Walchand College of Engineering, Sangli*
3	Dr. S. H. Bhandari	*Head of Department & Associate Professor - Department of Computer Science & Engineering, Walchand College of Engineering, Sangli*
4	Dr. D. B. Kulkarni	*Professor - Department of Information Technology, Walchand College of Engineering, Sangli*
5	Dr. S. P. Sonawane	*Head of Department & Associate Professor, Department of Information Technology, Walchand College of Engineering, Sangli*
6	Mr. A. R. Surve	*Assistant Professor - Department of Computer Science & Engineering, Walchand College of Engineering, Sangli*
7	Mrs. M. A. Shah	*Assistant Professor - Department of Computer Science & Engineering, Walchand College of Engineering, Sangli*
8	Mr. M. K. Chavan	*Assistant Professor - Department of Computer Science & Engineering, Walchand College of Engineering, Sangli*

PREFACE

Dr. G. V. Parishwad
Director, WCE Sangli

Dr. P. J. Kulkarni
Dy. Director, WCE Sangli

In Higher and Technical Education system, faculty members with higher research qualifications, especially in Information Communication and Technology (ICT) area, have to accept an important challenge of producing more number of researchers who in long run will take up responsibility of educating masses on design and use of modern technological gadgets for betterment of mankind. Over the last two decades, the growth of ICT products has been phenomenal. Rate of obsolescence of these ICT products is also significant. In order to cope with advancement of Engineering and Technology, researchers in the educational institutes need to brain storm, exchange their ideas with peer groups and make their research more fruitful. In line with this theme, Computer Science & Engineering Department and Information Technology Department of Walchand College of Engineering (WCE) proposed for holding of symposium in Computing. With constant proactive efforts of faculty members of these departments, the symposium scheduled in December 2016 has taken up a good shape for its organization. WCE has also been recognized to conduct Ph.D. research programme under Quality Improvement Programme (QIP) of AICTE. Along with this, Technical Education Quality Improvement Programme (TEQIP), being implemented in its phase II at WCE has one of the important components of promoting research culture among the budding engineers and faculty members. Therefore the organization of the symposium attracted a very good support from TEQIP. With very good response received from researchers for their paper presentations and with confirm availability of very good practicing researchers for delivery of key note speeches at the symposium, we are confident that the organization of the symposium will be a grand success.

We would like to encourage all the participants of this symposium to take positive and active part in the deliberations at the symposium. On behalf of the college management, we wish them bright and fruitful career ahead.

Director **Deputy Director**

Walchand College of Engineering, Sangli

From the desk of the Chair RSC 2016

Dr. P. J. Kulkarni
Dy. Director and Professor in CSE
The Chair RSC 2016

Search and Research are continuous activities. Research culture in Computer Science and Engineering (CSE) at Walchand College of engineering (WCE), over last decade has seen significant positive growth. Year after year, more research outcomes are being strengthened. Quality Improvement Programme (QIP) of AICTE has instituted Ph D research center in CSE at WCE. Along with this, Shivaji University Kolhapur has already recognized the CSE department to conduct Ph D research programme. Association of Computing Machinery (ACM) also identified the CSE department to institute student chapter of ACM. The chapter activities are well progressing. To further encourage CSE research activities, it is envisaged to create a platform to enable researchers in the field of CSE and allied to come together to provide critique on the ongoing research activities to enable shape these activities in a better way. In this direction, CSE Department and Information Technology (IT) department at WCE decided to jointly organize a "National Research Symposium on Computing, RSC-2016". The student chapter of ACM at WCE came forward to support the organization. At WCE, this is first of its kind of organization of the research symposium. Right from the day of its announcement, few months back, the organizing team of the conference started receiving very good responses from research community in CSE and IT. The research papers selected for the symposium are duly peer reviewed by outside research experts in the respective domains. The paper reviewing experts have technically well contributed by providing prompt and critical inputs to the authors of the papers. In order to provide good mentoring to the young researchers and attendees, the organizers of RSC-2016 are fortunate to attract good number of practicing researchers to deliver key note addresses. A pool of expert panel members will provide to the attendees of the symposium a very good exposure to state-of-art in CSE, IT and allied fields. The entire focus of the symposium is to facilitate budding researchers to bring in innovations in their on-going research and make the research fruitful. Many of the participating researchers in the symposium are believed to have come from academic institutes, therefore various issues related to good practices in research methodologies, peer-to-peer sharing, widening contacts of like-minded researchers, presentations of research work, Intellectual Property Rights (IPR) etc. will be well deliberated. I am very much confident that the symposium will mark significant achievement for all attendees of the symposium in practicing quality research work. I wish each one of them excellent prospective research career in future.

Dr P J Kulkarni, Chair, The RSC 2016

Key Note Speech – I

From ACM and IEEE distinguished speaker

Towards Awakening Research Whereabouts

Aditya Abhyankar, *Senior Member, IEEE*

Abstract: Research has become a buzz word in India in last decade or so. While major part of it comes in the form of compulsion by some authoritative body, it is the choice of an individual to maintain the quality of the research. There are connotations associated with research whereabouts. This presentation is an effort to deliberate on the quality norms with the research fraternity and evaluate the parameters to form some consensus in the minds of researchers. The journey from abstraction to creating tangible forms will also be discussed. The importance of IP (Intellectual Property) generation and the dissemination of the same for commercialization in some form is one area often ignored by academicians. This talk will also try and help academicians understand the importance of creation of eco-system and help budding researchers leverage the same to come up with sustainable model of good quality research.

Index Terms: **Research, Intellectual Property, Quality of Research, Commercialization, Plagiarism, Quality Norms.**

I. INTRODUCTION

THE moment we start discussing research, the quality norms play important role and they should be carefully evaluated and understood to make ones research stand out and count.

The criticalities in the form of compulsions from various bodies of authority have pushed academicians in the realms of research. The entire journey of research from abstraction to tangible forms is interesting as well as challenging. Though what gets achieved from the research effort is largely dependent on the ability of the researcher, there are certain quality norms which can be accepted in universal form. It is important to ask the question why research!

II. RESEARCH WHEREABOUTS

A. Important characteristics of good quality research:

1. The work should be replicable and reproducible
2. The work should be generalizable
3. The work should be based on sound rationale
4. The work should be incremental
5. The work should have good societal impact

B. Important characteristics of bad research

1. Plagiarizing other work
2. Falsification of data
3. Misrepresentation of the data and dubious conclusions

C. Important Research objectives

1. To solve interesting complex problems

2. To make new discoveries
3. To help develop new products
4. To optimize existing systems and bring betterment in the workflows
5. To help satisfy human desires and curiosity
6. To investigate laws of nature
7. To earn higher degrees (M.Phil / PhD etc)

Good research always begins with a good question or enquiry. It is then important to convert good research into a technical document, either a paper or a thesis.

III. POPULAR MISCONCEPTIONS

1. Technical paper should have lot of maths
2. The article should be hard to read
3. One must fill up all the pages

IV. PATENTING BAREBONES

To be able to patent the idea should be

1. Novel
2. Non-obvious
3. Useful

V. CONCLUSION

Research is addictive and can give great satisfaction; it however requires perceived efforts, eye for detailing and smart choices. Most of the inventions are apparent accidents and the accidents are bound to happen, but one has to walk the path first. Through such interactions, hopefully, the readers will start finding the paths and will start walking the same, which eventually will result into interesting technical accidents!

About Speaker: **Dr. Aditya Abhyankar**

Dr. Aditya Abhyankar is currently working as Dean, Faculty of Technology and Professor in Department of Technology of SP Pune University. He is on lien from his duties as Dean R&D, Director CERD (Center for Excellence in R&D) and Professor in Computer Engineering Department of VIIT, Pune. He is associated as an adjunct professor with COEP, Pune, as research associate with Clarkson University, NY, USA and on advisory committee and BOS (board of studies) of many national and international universities.

Aditya Abhyankar received the BE degree in Electronics and Telecommunication Engineering from Pune University, India in 2001. He received the MS and Ph.D. degrees from Clarkson University, NY, USA in 2003 and 2006 respectively. He worked as a post-doctoral fellow at Clarkson University, NY, USA in the academic year 2006-07. He worked as a consultant for Biometrics LLC, WV, USA in 2007. Aditya has also earned M.A. (Sanskrit) with specialization in Vedant where he stood first in university in 2011, M.B.A. with specialization in finance in 2013 and M.A. (Philosophy) with specialization in Existentialism in 2015. He is pursuing his second

Ph.D. in Sanskrit with his research focusing on Nasadiya sukta from Rigved. His research interests include pattern recognition, signal and image processing, wavelet analysis and biometric systems. Dr. Abhyankar has also won number of national and state awards with the likes of IEI scientist of the year 2011 award, sir C V Raman 2015 award at the hands of Hon'ble CM of Maharashtra to name a few. He has contributed in national level programs for education like NPTEL at IIT Bombay. Dr. Abhyankar holds 8 US patents, 14 Indian patents, 7 disclosures, 8 Technology Transfers, 5 US copyrights to his credit. He has more than 70 international journal articles, 180 international conference papers. He has generated research funding of the tune of 2.5 million USD and has created state-of-the-art infrastructure in the field of pattern recognition. He provides consultancy to number of reputed industries like Biometrics LLC, USA, vision&D, Canada, Optra Systems, India, BYDesign, Banglore, India etc. He also is director of Adijit SofTech Pvt Ltd which works in creating intelligent frameworks for pattern recognition based solutions.

Key Note Speech – 2

From Research Innovation Lab Speaker

Knowledge Innovation and Machine Learning

Parag Kulkarni
CEO and Chief Scientist, Iknowlation Research Labs Pvt Ltd Pune
(www.iknowlation.com)

Traditionally it was more about knowledge acquisition based learning. Knowledge acquisition based learning came up with wonderful methods right from statistical regression based methods, SVM, Bayesians to Deep and reinforcement learning. Reinforcement Machine Learning advocated exploration, which can be termed as dynamic knowledge acquisition. Dynamic Knowledge Acquisition set a new trend in Machine Learning (ML). While traditional mapping based learning demands data, too much context and restricts the learnability, reinforcement tried to introduce the ability to explore. We measure accuracies but hardly learnability of algorithm. There are limitations of Knowledge Acquisition based learning paradigm. With evolution of system and applications demanding more we need to move to creative and rather knowledge innovation based ML. This presentation focuses on Knowledge Innovation based Machine Learning catering to real life applications of Machine Learning. With numerous interesting case studies the aspects of creative machine learning will be discussed. All traditional methods right from rule based systems to recent advanced methods approach learning from same paradigm. As mentioned by Margaret Boden – can we build really creative machines? And if we want to build these machine will these learning paradigms are suffice. Well we need different paradigms – we need to look and solve the learning problems in different way where learnability will be at centre. While taking crack at this learning problem, the talk will try to touch on different real life ML case studies. It will introduce some pioneering research in machine learning by iknowlation including Optimal Machine Learning, DEEP reinforcement Machine Learning and Context Vector Machines. In this presentation, I will argue that Optimal Creative ML is going to be next trend in machine learning. Over learning is dangerous so is under learning. While over learning kills creativity – under learning leaves us without clues. It is the time to go for Optimal Machine Learning. Find out more about pointers to optimal learning during this presentation that plans to cover landscape of opportunities in ML with focus on real life applications and research directions.

About Speaker: **Parag Kulkarni**

Parag is an entrepreneur, Machine Learning researcher and author of best selling Innovation Strategy and Data science books. An avid reader, Parag is founding CEO and Chief Scientist of iknowlation Research Labs – a vibrant Machine Learning Product, research and Consulting Company. Parag has published over 300 research papers, invented over a dozen patents and he authored 15 books. Parag's machine learning ideas resulted in pioneering products those became commercially successful and produced unprecedented impact. As a consultant Parag has contributed to success of over two-dozen organizations including start-ups and established companies. He is pioneer of concepts Systemic Machine

Learning, Context Vector Machines and Deep Explorative Machine Learning. He delivered over 300+ keynote addresses and 200+ tutorials across the globe. An alumnus of WCE Sangli, Parag holds PhD from IIT, management education from IIM and was conferred higher doctorate DSc by UGSM monarch, Switzerland. His work on Systemic Machine Learning published by IEEE is widely cited. His areas of interest include Machine learning and allied areas with focus on optimal and systemic learning.

Key Note Speech – 3

From The IET and Practicing Consultant Speaker

Three-fold Approach to Innovation

Lokesh Venkataswamy
Managing Director
Innomantra Consulting Private Limited, Bangalore
Lokeshv@innomantra.com

Abstract: *Innomantra is India's leading Innovation and Intellectual property consulting and services firm. It helps organisations to design and achieve their innovation and Intellectual property goals by 3x. Its clients range from small entrepreneurial enterprises to Fortune Global 500 organisations. The firm is headquartered in Bengaluru, India. Innovation3x describes the firm's philosophy which believes that innovative organisations must identify innovation goals that seeks to achieve a 3x boost in performance. It also represents three-fold approach to innovation that looks at overall business strategy, people and functional systems using Innovation Maturity Model, Collaborative Innovation, Design Thinking and Functional Innovation Methodologies.*

About Speaker**: Lokesh Venkataswamy**

Lokesh is Managing Director at Innomantra. He brings with him more than 20 years of professional experience in Intellectual Property management, New Product Development, and innovation management. Prior to Innomantra, he managed innovation and intellectual property activities at Larsen & Toubro. He began his career with the Medical Equipment and Systems Group at L&T where he spent 12 years in product design & development, strategic planning & initiatives driving Innovation & IP. At Innomantra, Lokesh has worked with clients from several Fortune Global 500 to SME's in a wide range of sectors including Automotive, Industrial Equipment Manufacturing, Information Technology, Engineering Services, Medical Electronics and Consumer Goods in the areas of Intellectual Property Management and Innovation Management. Lokesh, a Mechanical Engineer, has an MBA in Marketing and PG Diploma in Intellectual Property Law from National Law School of India University (NLSIU). Lokesh was also awarded winner of "India Innovation Challenge" at 4th India Innovation Summit - 2008 Organized by CII & Govt. of Karnataka. Lokesh also played a role of mentor and jury in 'Power of Ideas 2010,2012 and 2015– Initiated by top business daily in India 'The Economic Times' and Indian Institute of Management –Ahmadabad, India' and supports National Entrepreneur Network(NEN) India. He is also a member and Chairman – Bangalore Local Network, Institution of Engineering & Technology (IET) and Member at IT Panel at CII, India. He is also a Co-Founder & Trustee – Association of Intellectual Property Professionals & Owners Association (AIPPO), India.

Key Note Speech – 4

From The IET Speaker

Data Science Vs. Big Data for Business Decisions

Sunita Patil

Vice Principal & Dean Academics & Professor, Department of Computer Engg.

K J Somaiya Institute of Engg. & IT Mumbai

Data have become a real resource of interest worldwide for every industry & business. Along with the rise of data two distinct efforts concerned with harnessing its potential, one is called Data Science and the other as Big Data. These terms are often used interchangeably despite having fundamentally different roles to play in bringing the potential of data of an organization.

Big data is a term that describes the large volume of data; both structured and unstructured that covers a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters.

Data science is an interdisciplinary field about scientific processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics, similar to Knowledge Discovery in Databases (KDD).

Data Science and Big Data both are related to Data Driven Decision with the expectations of better decision and increased value. This is a process and involves different stages as Capture the Data, Process & Store it, Analyse and Generate Insights into data and then take Decisions & Actions. In this data driven decision process, Big Data is typically involved in second stage and that too in all the technologies. Big Data & Technology helps in reducing cost in processing volume of data and also making it feasible to do a few typically analyses whereas Data Science is involved at the third stage in data driven decision process. It involves in using Statistical, Mathematical and Machine Learning algorithms to use data and generate insights. Whether a data is "Big data" or not, we can use Data Science to support Data Driven Decisions and take better decisions.

Data Science looks to create models that capture the underlying patterns of complex systems, and codify those models into working applications while Big Data looks to collect and manage large amounts of varied data to serve large-scale web applications and vast sensor networks.

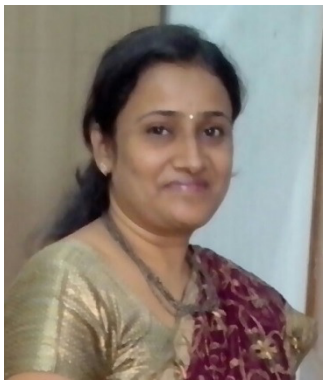
Although both offer the potential to produce value from data, the fundamental difference between Data Science and Big Data can be summarized as collecting data does not mean discovering potential in it.

In the world of data science, it takes a science to convert a raw resource into something of value is because what is extracted from the 'ground' is never in a useful form. 'Data in the raw' has noise, irrelevant information, and misleading patterns. To convert this into precious thing requires a study of its properties and the discovery of a working model that captures the behavior we are interested in. Something unique to business that has given them the knowledge of what to look for, and the codified descriptions of a world that can be mechanized and scaled as further discovery and innovation. No organization would invest in the extraction of a resource without the expertise in place to turn that resource into value.

Unfortunately with Big Data, the largest companies see what solutions they have engineered to compete in their markets. But these companies hardly represent the challenges faced by most organizations. Their dominance often means they face very different competition and their engineering is done predominantly to serve large-scale applications. This engineering is critical for daily operations, and answering to the demands of high throughput and fault-tolerant architectures. But it says very little about the ability to discover and convert what is collected into valuable models that capture the driving forces behind how their markets operate. The ability to explain and predict an organization's dynamic environment is what it means to compete using data.

Understanding the distinction between Data Science and Big Data is critical to investing in a data strategy. For organizations looking to utilize their data as a competitive asset, the initial investment should be focused on converting data into value. The focus should be on the Data Science needed to build models that move data from raw to relevant. With time, Big Data approaches can work in concert with Data Science. The increased variety of data extracted can help make new discoveries or improve an existing model's ability to predict or classify.

About Speaker**: Sunita Patil**



Sunita R. Patil completed her graduation from North Maharashtra University, Jalgaon in the year 2000 and post-graduation from University of Mumbai in the year 2008 and is currently working as Vice Principal & Dean Academics & Professor,Department of Computer Engg. at K J Somaiya Institute of Engg. & IT Mumbai. She completed her PhD in Data Mining from NMIMS, Vile Parle, Mumbai. She is the convener for IET (Institute of Engineering & Technology, Students Chapter & Academic Affiliation). She is the Vice Chairman for IET Mumbai Local Network & Ex-VC Young Professional of IET Mumbai Local Network. She has specialization in various fields like Data Warehousing Mining & Business Intelligence. She has authorized more than 16 research papers at international, national journals and conference proceedings.

Key Note Speech – 5

From The University Research Group Speaker

Internet of things: A Novel Interdisciplinary Research Paradigm in Computer Science

Professor (Dr.) R.K. Kamat
Head, Department of Electronics,
Shivaji University,
Kolhapur – 416 004

Abstract: *A logical extension of the embedded and/or cyber physical systems is the conception of Internet of Things (IoT). Popularly known as the 'Internet of Everything' this novel paradigm is now growing in leaps and bounds and is almost about to touch each and every walks of societal life.*

The IoT based systems are characterized by different subsector components such as the sensors, instrumentation, communication protocols, different means of attaining the connectivity, smartphone apps and simultaneously making the system work in the constraint of power, bandwidth as well as other spatial and temporal metrics.

The present talk will exemplify various IoT based systems, their design and deployment in the application domain. The talk will emphasize on the development lifecycle of such systems and the way the interdisciplinary thinking amongst the various branches of Engineering, paving the ways for their development.

About Speaker**: Dr. R. K. Kamat**

Professor R.K. Kamat is currently Professor and Head of the Department of Electronics, Shivaji University, Kolhapur. He has extensive teaching and research experience spanning over last two decades. He has published 100+ papers in the international journal of repute. Eleven students have been awarded Ph.D. degree under his guidance. He is also recipient of the Young Scientist award of the Department of Science and Technology (DST), Government of India. Currently he is also Vice-President of the IEEE SSCS, India Council. Government of Maharashtra has recently appointed him on the State Level Quality Council (SLQC) which is mandated the quality inculcation in the institutes of Higher Education in the state of Maharashtra.

Key Note Speech – 6

From Invited Speaker

Mosaicing Scenes with Vacant Spaces through a Quadcopter

Meghshyam G. Prasad CSE
Department, IIT Bombay
meghshyam@cse.iitb.ac.in

Abstract—This paper focuses on a method of constructing panoramas from a quadcopter, and a new mosaicing sub-problem when the scene contains significant regions of vacant spaces. These vacant spaces yield little to no features to match input images and hence challenge existing mosaicing techniques. We describe a framework that is able to handle this unique input by leveraging the availability of the inertial measurement unit (IMU) data from the quadcopter. We demonstrate the efficacy of our approach on a number of input sequences that cannot be mosaiced by existing methods.

I. Introduction

Finding features and using them to align images to construct panoramas is one of the success stories of computer vision. Virtually all recent consumer cameras have this technology embedded. The success of these methods relies significantly on finding common features in the images that can be used to establish the appropriate warps to register the images together.

Problem Definition: There are scenes, however, that makes this challenging. One situation is when the scene needs to be probed in an orthographic view, and is not easily accessible. Murals on large urban architecture is an example. This suggests a 'close up orthographic view' by a moving camera fitted on, say a quadcopter. A related situation is when a scene area simply does not contain features (too little, or no features, e.g. multiple adjoining posters in an conference event).

The goal of this paper is to compute a panorama of a scene lying on planar surface that is, possibly, difficult to reach manually. The scene is assumed to have vacant spaces. A schematic for this problem is shown in Figure 1.

II. Related Work

Panoramic image stitching (alternatively, image mosaicing) is a well-studied problem in the field of computer vision. Representative works include [1], [2], [3], [4] [5] [6]. A full discussion on related works is outside the scope of this paper, readers are referred to [7] for an excellent survey. Given the maturity of this area, there are various freeware as well as commercial software available for performing image stitching; most notable are AutoStitch [8], Microsoft´s Image Compositing Editor [9], and Adobe´s Photoshop [10].

All of these methods are based on a similar strategy of finding features in each image, matching these features between images, and then computing pairwise image warps to align them together. A bundle adjustment is often applied to globally refine the alignment. All of the aforementioned methods assume the imaged scene is planar or that the camera has been rotated carefully around its center of projection to avoid parallax.

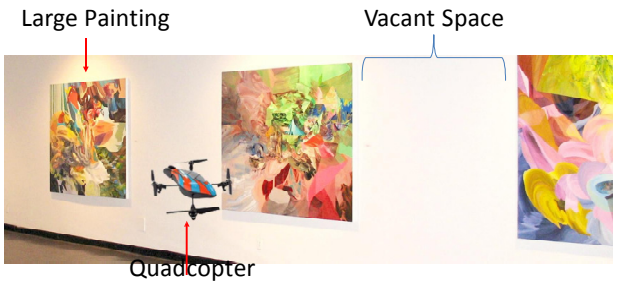


Fig. 1. Problem definition. Vacant spaces are encountered in various scenes. When individual portions are captured by a quadcopter, how does one create the complete mosaic given that common features are either not available as in this example, or confusing (see Fig. 7)?

Brown et al. [11] proposed a method that used invariant features located at Harris corners in discrete scale-space and oriented using a blurred local gradient for stitching. Eden et al. [12] were able to stitch images with large exposure difference as well as large scene motion into single HDR quality image without using any additional camera hardware.

All of the image mosaicing methods work only when there is an "intersection" in feature space of images to be stitched. When there are "gaps" (either physical or due to lack of features) between images to be stitched it is not clear how to perform the stitching.

III. Methodology

The method adopted is pictorially depicted in the overview shown in Figure 2 and is described in detail later on. In brief, we systematically acquire a video of the scene, reduce the input video to a manageable number of images, and finally combine the images acquired from different positions into a mosaic.

A. Video acquisition

We first dispatch the quadcopter to as close to the scene as possible. The corners of a rectangular area of interest are provided to the quadcopter, and it is programmed to traverse

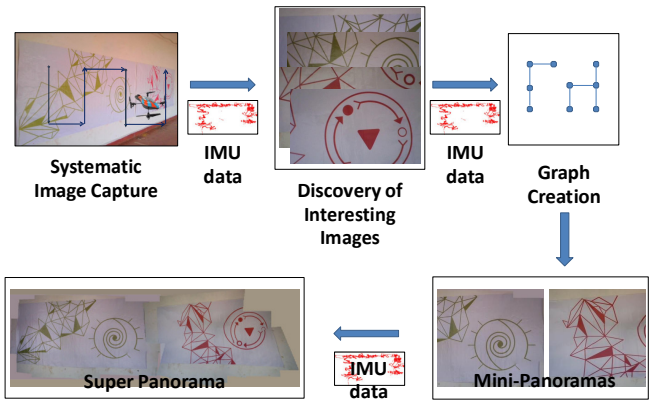


Fig. 2. Overview: Input imagery is systematically acquired (top left) by a quadcopter. In the next step, interesting images are found by clustering the video into regions based on positional data. A graph is constructed using proximal images. For each connected component in a graph, standard stitching techniques are used to create mini-panoramas which are then joined together into super panorama again using the IMU data.

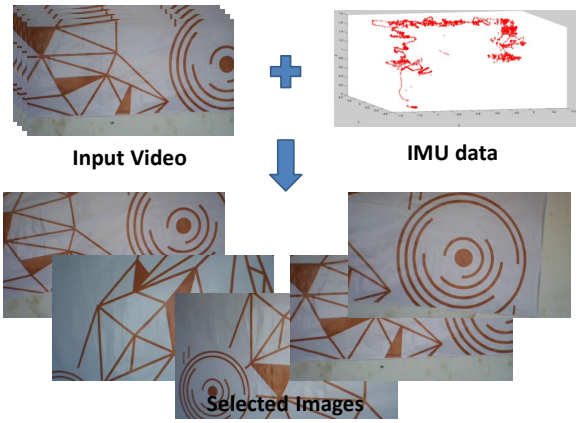


Fig. 3. We align the image stream with the IMU data, and then transform the video into a set of interesting images with a clustering algorithm.

the area in a raster scan fashion. Note that trying to create a mosaic in an incremental linear fashion by combining adjacent frames is prone to loss of two-dimensional spatial proximal information. It is also computationally overwhelming.

The quadcopter returns with a video of the scene. Images extracted from a short video of about a minute or more overwhelms existing mosaicing software, such as AutoStitch or Adobe Photoshop.

B. Selecting interesting images

Our goal in this step is to reduce the amount of input data and produce a set of interesting images. In other words, we wish to convert a video into an album of images. The key difference between our problem and standard albumization [13], [14] is the use of positional information. A standard quadcopter has an IMU that, after calibration, may give reasonable information of positions. Using positional information it is possible to cluster the images, and sort the images into an $m \times n$ grid. The number of cluster centers is automatically determined using the agglomerative bottom up hierarchical clustering method [15], with the additional requirement that the whole scene (represented by the positional data) is covered.

Clustering Details We assume that each IMU data position corresponds to an image of definite fixed dimensions. Consider each position of the IMU data to be a leaf node. Two nodes are greedily combined based on the closest Euclidean distance, and replaced with an internal node; the position of the internal node is set to be the centroid of the two nodes, and each internal node now corresponds to a virtual image of the same size taken by a virtual quadcopter. The algorithm recursively merges all the nodes till we end up with a root. In the next phase, we produce cluster centers; a set of nodes is considered for being the output as cluster centers if the union of these nodes completely spatially cover the scene. From the bottom-up construction, it is clear that the root will represent a single position, and thus a single virtual image, and will not cover the scene. At the other extreme, the set of all leaf nodes will cover the scene. To resolve this, during the calibration phase, we pre-decide the minimum distance between two center of projection to have least overlap. This is used as the threshold in the clustering algorithm. Once cluster centers are found, we pick the leaf node which is closest to the cluster center to find a real image. This process is schematically shown in Figure 3.

In practice, the number of cluster centers and thus images for the scenes we have covered is now within the capacity of AutoStitch. As mentioned in the introduction, as long as there are sufficiently varying and "matchable" features, AutoStitch is able to perform a reasonable result. However, if there are very few features in overlapping region of two images, then the output is not acceptable. This situation will arise when there is vacant space in the imagery.

Mini-Panoramas Specifically, we assume at this point that the interesting photos are available in the form of a $m \times n$ grid. First, we find SURF [16] features for each image in a grid. Next, we use Best of Nearest Neighbor matcher (from the OpenCV library) with Random Sample Consensus (RANSAC) [17] to find geometrically consistent matches between neighborhood images inside grid. We create a graph with images being nodes, and add an edge between two nodes if there are sufficient matches. We have to recall at this point that if there are "vacant spaces" there will not be enough features for successful matches; the graph will end up with multiple (disconnected) components. We next compute multiple spanning trees for the various components. Given a spanning tree, the center of the spanning tree is a node from which the distance to all other nodes is minimal [18]. Next we calculate the homography of each image with respect to spanning tree center. Finally, for each spanning tree, we stitch all pictures within the spanning tree to create a mini-panorama using the computed homographies by warping all images with reference to the image at spanning tree center. The spanning tree is an O(N) structure. The process is described in Figure 4.

C. Super-panorama

In this section, we consider the situation when programs like AutoStitch fail. We assume that the output of the previ-

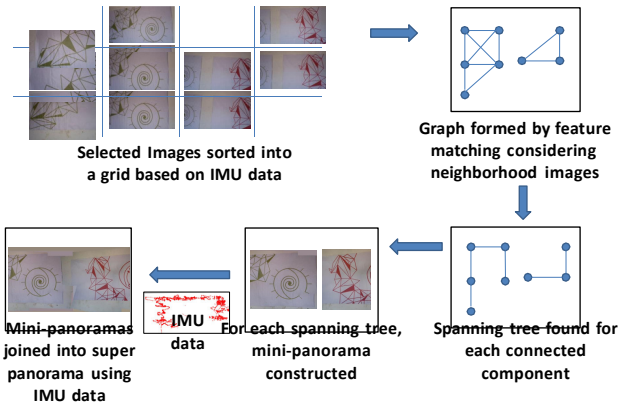


Fig. 4. Interesting images acquired are segmented and individual (mini-panoramas) are constructed. These are then later combined into the desired super-panoramas using the IMU data.

ous step has resulted in multiple spanning trees where each spanning tree center corresponds to a specific depth, since we have stitched all images by taking the spanning tree center as a reference. Individual panoramas for each spanning tree termed mini-panoramas have been created. A super-panorama must be created from mini-panoramas; these usually correspond to different depths due to jerky motion of quadcopter.

A super-panorama is done using a two step process. Assume two trees in the forest corresponding to area A and area B of the scene (see Figure 5). Assume that a mini-panorama is created from these two areas, and the depth of the planar surface from the camera is more for A, than for B. We then take the mini-panorama image captured at B, and 'move' it to a new location B' whose depth (from the imaged surface) is the same as that of A. The resulting image will be smaller; the images are related by the equation

$$\frac{x'}{x} = \frac{Z}{Z'} \tag{1}$$

where x (respectively x') represents a pixel location of the image in B (respectively, B') and Z (respectively Z') represents the depth of the images surface from B (respectively, B').

In order to form a super-panorama from the depth of A, we can now treat the resulting images from A (unchanged) and B (computed from Equation 1) forming a simplistic stereo pair at the depth of position A. Using the stereo disparity formula we can "place," from the view point of A, the image captured from B', thereby creating a super-panorama. (We could as well present the entire scene from the viewpoint of B' (since it is at the same depth); we prefer these pictures to the one that one may be created from the depth of B.)

IV. Experiments and Results

All our experiments have been completed with the inexpensive consumer quadcopter called a Parrot's AR Drone 2.0. We have implemented our algorithm in C++ using the OpenCV library. Experiments were performed on a PC with

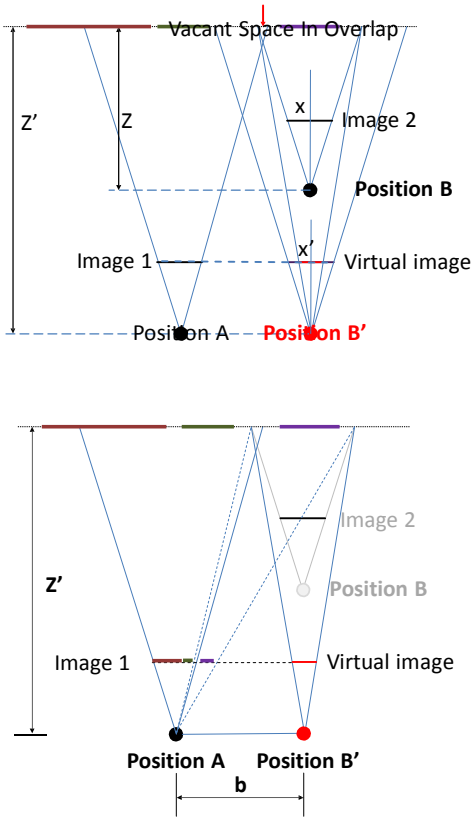


Fig. 5. (Top) The virtual picture as seen from position B' is computed using Equation 1 from the real picture taken from B. (Bottom) Using the stereo disparity, calculated from the baseline width b and depth Z', it is possible to depict the composite scene obtained from both A and B' (from the view point of A).

Intel Core i7 processor(@3.4GHz) and 8GB RAM. Please visit http://goo.gl/sYvoVP for datasets and code related to the paper.

A. Indoor Imagery with Vacant Spaces

Our first set of experiments were conducted in an indoor environment.

The input stream had about 9000 input images. The selection algorithm pruned the video into N = 13 images. A sample of the selected images are seen in Figure 6(a). The scene as captured by a smartphone can also be seen, as well as the outputs of the state of the art stitchers. Note that AutoStitch is only able to stitch the upper half of the scene. Our result Figure 6(e) clearly stands out in comparison.

B. Outdoor Imagery with Vacant Spaces

Our next set of experiments were conducted in an outdoor environment. The input stream had about 12000 images. The selection algorithm pruned the video into N = 30 images. A sample of the selected images are seen in Figure 7(a). The scene as captured by a smartphone can be seen in Figure 7(b). Figures 7(c), (d) and (e) shows the comparison of outputs

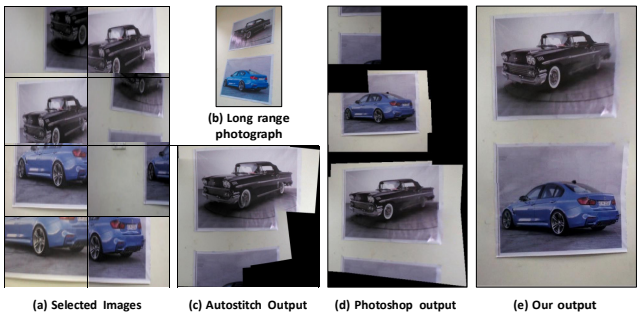


Fig. 6. (a) Salient images from the quadcopter video using our selection algorithm of (b) an indoor scene. This long range photograph has been captured separately by a smartphone camera only for context. Notice a significant vertical vacant space (around three feet) in the imagery. (c) Output of AutoStitch – only the upper half of the scene is output. (d) Output of Adobe Photoshop CS6 – the vacant space posed a problem to the feature matching algorithm, so instead of a mosaic, individual pieces were output as mini-panoramas (e) Our output on the selected images. We are able to present the scene in high fidelity in an orthographic view.

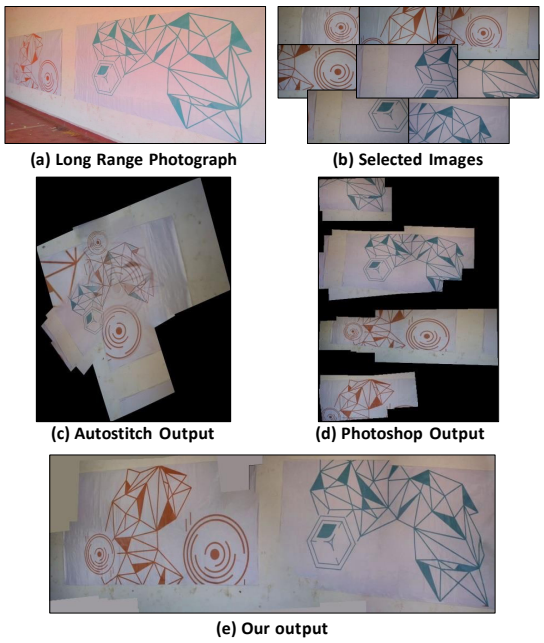


Fig. 7. (a) An outdoor scene captured by a standard camera in an exhibition. The approach to the area is normally cordoned off and one needs permission to get a quadcopter to take the picture. Notice a significant gap (more than two feet) between the two posters. (b) Salient images from the quadcopter video using our selection algorithm. (c) Output of AutoStitch on the selected images. The mosaic is not reasonable presumably because of the confusion in features. (d) Output of Adobe Photoshop CS6 on the selected images. The vacant space posed a problem to the feature matching algorithm, so instead of a mosaic, individual pieces were output as mini-panoramas (e) Our output on the selected images. We are able to join two posters (separated by vacant space) using the IMU data.

of state of the art stitchers with the output of our algorithm. Note that AutoStitch is getting confused by too many matching features.

V. Concluding Remarks

In this paper, we have described a method of imaging large scenes using a quadcopter enabling close orthographic views. We also defined a new problem, that of computing a mosaic of a planar scene with vacant spaces. Vacant space relates to images in an input stream where there are not enough features for traditional mosaicing algorithms to estimate geometric warps to align the images.

Our solution to this problem is to use an autonomous quadcopter which is capable of taking pictures. The quadcopter has an inertial measurement unit that is capable of outputting reasonable spatial locations, but unreliable roll, yaw and pitch. Using this positional information, our algorithm selects an "interesting" subset of the video imagery. Whenever there is overlap in feature space in the subset, we stitch images using traditional vision-based methods avoiding the erroneous roll based warping. At other times, we use positional information, and reduce the resulting problem of computing a complete panorama to the stereo problem of merging mini-panoramas. Our method works on both indoor and outdoor scenes.

References

[1] D. Milgram, "Computer methods for creating photomosaics," IEEE Transactions on Computers, vol. C-24, no. 11, pp. 1113–1119, 1975. 1

[2] ——, "Adaptive techniques for photomosaicking," IEEE Transactions on Computers, vol. 26, no. 11, pp. 1175–1180, 1977. 1

[3] D. Capel and A. Zisserman, "Automated mosaicing with super-resolution zoom," in CVPR, 1998. 1

[4] R. Szeliski and H.-Y. Shum, "Creating Full View Panoramic Image Mosaics and Environment Maps," in SIGGRAPH, ser. SIGGRAPH '97, 1997, pp. 251–258. 1

[5] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," International Journal of Computer Vision, vol. 74, no. 1, pp. 59–73, 2007. 1

[6] M. Brown and D. Lowe, "Recognising panoramas," in ICCV, 2003. 1

[7] R. Szeliski, "Image alignment and stitching: A tutorial," MSR-TR-2004-92, Microsoft Research, 2004, Tech. Rep., 2005. 1

[8] AutoStitch, http://matthewalunbrown.com/autostitch/autostitch.html, [Online; accessed 23-January-2016]. 1

[9] Microsoft Image Composite Editor, http://research.microsoft.com/en-us/um/redmond/groups/ivm/ICE/, [Online; accessed 23-January-2016]. 1

[10] Adobe Photoshop CS6, http://www.adobe.com/products/photoshop, [Online; accessed 23-January-2016]. 1

[11] M. Brown, R. Szeliski, and S. Winder, "Multi-image matching using multi-scale oriented patches," in CVPR, 2005. 1

[12] A. Eden, M. Uyttendaele, and R. Szeliski, "Seamless Image Stitching of Scenes with Large Motions and Exposure Differences," in CVPR, 2006. 1

[13] A. Aner and J. R. Kender, "Video summaries through mosaic-based shot and scene clustering," in ECCV, 2002. 2

[14] Y. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in CVPR, 2012. 2

[15] L. Rokach and O. Maimon, "Clustering methods," Data mining and knowledge discovery handbook, pp. 321–352, 2005. 2

[16] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," in ECCV, 2006. 2

[17] M. A. Fischler and R. C. Bolles, "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography," Commun. ACM, vol. 24, no. 6, pp. 381–395, Jun. 1981. 2

[18] W. Kocay and D. L. Kreher, Graphs, Algorithms, and Optimization, ser. Discrete Mathematics and Its Applications. Chapman and Hall, 2004. 2

Generating Informative Summary of Social Image Search Result

Sheetal. A. Takale
VPKBIET Baramati, India
SPPU, Pune, India
sheetaltakale@gmail.com

P J Kulkarni
Walchand College of Engineering, Sangli, Maharashtra, India
pjk_walchand@rediffmail.com

Abstract

Tag based social media search presents search results as a ranked list of images. However, the search results are not informative. Informative summary is with regards to the identification of: important visual concepts, locations-of-interest (LOI), and context describing tags. In this paper, we propose to generate an informative summary of social image search results. The proposed scheme exploits multiple modalities in order to understand context and content of geo-tagged social images. We formulate the problem as a graph clustering problem, where edge weights are geographical distance, textual and visual similarity between images. Late fusion of three different edge weight parameters avoids computational overhead. Performance evaluation is based on intrinsic and extrinsic methods using both Ground Truth and an evaluation protocol having no human intervention. Through empirical study, we demonstrate the effectiveness of our algorithm to generate informative summary which is relevant, representative and diverse.

REFERENCES

1. D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," Annals of Data Science, vol. 2, no. 2, pp. 165–193, 2015.
2. F. Alamdar and M. R. Keyvanpour, "Effective browsing of image search results via diversified visual summarization by clustering and refining clusters," Signal, Image and Video Processing, vol. 8, no. 4, pp. 699– 721, 2014.
3. L. Cao, J. Luo, A. C. Gallagher, X. Jin, J. Han, and T. S. Huang, "A worldwide tourism recommendation system based on geotaggedweb photos," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, Sheraton Dallas Hotel, Dallas, Texas, USA, 2010, pp. 2274–2277.
4. M. Naaman, Y. J. Song, A. Paepcke, and H. Garcia-Molina, "Automatic organization for digital photographs with geographic coordinates," in In Proceedings ACM/IEEE Joint Conference on Digital Libraries, JCDL, Tucson, AZ, USA, 2004, pp. 53–62.
5. S. Wang, F. Jing, J. He, Q. Du, and L. Zhang, "IGroup: presenting web image search results in semantic clusters," in Proceedings of the 25th ACM SIGCHI Conference on Human Factors in Computing Systems. ACM, 2007, pp. 587–596.
6. B. Q. Truong, A. Sun, and S. S. Bhowmick, "Casis: A system for concept-aware social image search," in Proceedings of the 21st International Conference on World Wide Web, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 425–428.
7. Q. Hao, R. Cai, X. Wang, J. Yang, Y. Pang, and L. Zhang, "Generating location overviews with images and tags by mining user-generated travelogues," in Proceedings of the 17th International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, 2009, pp. 801– 804.
8. U. F. Niaz and B. M´erialdo, "Fusion methods for multi-modal indexing of web data," in 14th International Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS 2013, Paris, France, July 3-5, 2013, 2013, pp. 1–4.
9. H. Xu, J. Wang, X.-S. Hua, and S. Li, "Hybrid image summarization," in Proceedings of the 19th ACM International Conference on Multimedia, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 1217–1220.
10. B. Seah, S. S. Bhowmick, and A. Sun, "PRISM: concept-preserving summarization of top-k social image search results," PVLDB, vol. 8, no. 12, pp. 1868–1879, 2015.
11. L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, "How flickr helps us make sense of the world: Context and content in community-contributed media collections," in In proceedings, Fifteenth ACM International Conference on Multimedia (ACM MM 2007), Augsburg, Germany, 2007.
12. S. Rudinac, A. Hanjalic, and M. Larson, "Generating visual summaries of geographic areas using community-contributed images," IEEE Transactions on Multimedia, vol. 15, no. 4, pp. 921–932, 2013.
13. S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," IEEE Transactions on Multimedia, vol. 15, no. 6, pp. 1231–1243, 2013.
14. L. Kennedy and M. Naaman, "Generating diverse and representative image search results for landmarks," in In Proceedings, Seventeenth International World Wide Web Conference (WWW 2008), Beijing, China, 2008.
15. I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, "SUMMAC: a text summarization evaluation," Natural Language Engineering, vol. 8, no. 1, pp. 43–68, 2002.

Buddy Router : Novel DTN Routing Algorithm using Multiparameter Composite Metric

Ajit S. Patil,
PhD Research Scholar,
Dept. of Computer Science and Engineering,
Walchand College of Engineering, Sangli, Maharashtra, India
ajitspatil@gmail.com

Prakash J Kulkarni,
Dept. of Computer Science and Engineering
Walchand College of Engineering, Sangli
Maharashtra, India
pjk_walchand@rediffmail.com

Abstract

Delay Tolerant Networks (DTN) are special networks where nodes are thinly distributed and therefore long period of disconnections are observed in it. These sporadic connections, along with limited resources at mobile nodes pose challenging environment for routing in DTN. Opportunistic forwarding mechanism is explored by DTN routing strategies to deal with inherent transmission uncertainties. Maximizing message delivery ratio and reducing delivery overheads at the same time is the major objective behind number of opportunistic forwarding policies. This paper introduces two forwarding strategies: Buddy Router and its variant Buddy Router with Time Window. These are novel routing schemes for DTNs, which use new sociality metric called buddy metric. This buddy metric value is calculated by each node with every other node it has encountered with, using contact frequency, aggregate contact duration and contact recency between them. The main contribution of this paper is to present detailed formulation of Buddy Router and exploit buddy metric concept for efficient DTN routing. We also present comparative analysis of proposed routing strategies with existing DTN routing protocols.

Keywords—Delay Tolerant Network(DTN), Routing, Opportunistic Routing and Pocket Switched Networks (PSN)

REFERENCES

[1] K. Fall, “A Delay-Tolerant Network Architecture for Challenged Internets,” Intel Research Berkley, 2003. https://irtf.org/dtnrg

[2] Maurice J. Khabbaz, Chadi M. Assi, and Wissam F. Fawaz, “Disruption-Tolerant Networking: A Comprehensive Survey on Recent Developments and Persisting Challenges” IEEE Communications Surveys & Tutorials, Vol. 14, No. 2, Second Quarter 2012

[3] Yue Cao and Zhili Sun, Member, IEEE “Routing in Delay/Disruption Tolerant Networks: A Taxonomy, Survey and Challenges” IEEE Communications Surveys & Tutorials, Accepted For Publication.

[4] R. J. D’Souza, Johny Jose,NIT Surathkal, “Routing Approaches in Delay Tolerant Networks: A Survey” 2010 International Journal of Computer Applications (0975 - 8887)

[5] Artemios G. Voyiatzis, Member, IEEE, “A Survey of Delay- and Disruption-Tolerant Networking Applications” JOURNAL of Internet Engineering, vol. 5, no. 1, June 2012

[6] Ying Zhu, Bin Xu , Xinghua Shi, and Yu Wang “A Survey of Social-Based Routing in Delay Tolerant Networks: Positive and Negative Social Effects” IEEE Communications Surveys & Tutorials, Vol. 15, No. 1, First Quarter 2013

[7] Kaimin Wei, Xiao Liang, and Ke Xu, “A Survey of Social-Aware Routing Protocols in Delay Tolerant Networks:Applications, Taxonomy and Design-Related Issues” IEEE Communications Surveys & Tutorials, Accepted For Publication

[8] Paulo Rogerio Pereira, Augusto Casaca, Joel J. P. C. Rodrigues, Vasco N. G. J. Soares, Joan Triay, and Cristina Cervello-Pastor “From Delay-Tolerant Networks to Vehicular Delay-Tolerant Networks” IEEE Communications Surveys & Tutorials, Vol. 14, No. 4, Fourth Quarter 2012

[9] Amin Vahdat and David Becker "Epidemic Routing for Partially-Connected Ad Hoc Networks" Technical Report CS-200006, Duke University, April 2000.

[10] Lindgren, A. Doria "Probabilistic Routing Protocol for Intermittently Connected Networks" DTN Research Group, " ITRF 2012

[11] J. Lakkakorpi, M. Pitkanen, and J. Ott, " Adaptive Routing in Mobile Opportunistic Networks" ACM MSWiM 2010, Bodrum, Turkey, Oct. 2010, pp. 101-109

[12] P. Basu and S. Guha, "Effect of Limited Topology Knowledge on Opportunistic Forwarding in Ad Hoc Wireless Networks," Eighth International Symposium on Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WIOPT), Avignon, France, June 2010Buffer Management

[13] T. Spyropoulos, K. Psounis and C. S. Raghvendra "Spray and Wait Efficient routing in intermittently connected Networks," in Proceeding of Mobile Computer and Communication review Vol. 7,no. 3, July 2003.

[14] Burgess, J., Gallagher, B., Jensen, D., & Levine, B.N. (2006). MaxProp: Routing for Vehicle-based Disruption-Tolerant Networks. 25th IEEE International Conference on Computer Communications (INFOCOM 2006), 1-11.

[15] Henri Dubois-Ferriere , Matthias Grossglauser , Martin Vetterli, Age matters: efficient route discovery in mobile ad hoc networks using encounter ages, Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing, June 01-03, 2003, Annapolis, Maryland, USA [doi>10.1145/778415.778446]

[16] Pan Hui, Jon Crowcroft, and Eiko Yoneki, "BUBBLE Rap: Social-Based Forwarding in Delay-Tolerant Networks" IEEE Transactions on Mobile Computing, Vol. 10, No. 11, November 2011

[17] Eyuphan Bulut and Boleslaw K. Szymanski, "Exploiting Friendship Relations for Efficient Routing in Mobile Social Networks" IEEE Transactions On Parallel And Distributed Systems, Vol. 23, No. 12, December 2012

[18] Tamer Abdelkader, Kshirasagar Naik, Amiya Nayak, Nishith Goel, and Vineet Srivastava "SGBR: A Routing Protocol for Delay Tolerant Networks Using Social Grouping" IEEE Transactions On Parallel And Distributed Systems (Accepted for Final Publication)

[19] Shengling Wang And Min Liu, Xiuzhen Cheng, "Routing In Pocket Switched Networks" IEEE Wireless Communications, Feb 2012.

[20] Opportunistic Network Environment simulator. http://www.netlab.tkk.fi/tutkimus/dtn/theone/

[21] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau, "Data set cambridge/haggle," http://crawdad.cs.dartmouth.edu/ cambridge/haggle, may 2009.

[22] Pentland, R. Fletcher, and A. Hasson, "Daknet: Rethinking Connectivity In Developing Nations," Computer, vol. 37, no. 1, pp. 78 – 83, Jan. 2004.

[23] Mtibaa, M. May, C. Diot and M. Ammar "Peoplerank: Social opportunistic forwarding", IEEE INFOCOM '10, 2010

ABOUT AUTHORS

Mr. Ajit S. Patil received his degree in B.E. Computer Science and Engineering from Government College of Engineering, Aurangabad, India in the year 2000. He has completed his M. Tech in Computer Engineering from Dr. Babasaheb Ambedkar Marathwada University, Raigad, India. His M. Tech. dissertation work was based on, comparisons of different distributed token circulation algorithms in Mobile Ad hoc NETworks. His research interests include computer networks, mobile communications, and Delay Tolerant Networks (DTN). He is currently working as an Associate Professor in department of Information Technology at K.I.T.'S College of Engineering, Kolhapur, India and perusing his PhD at Walchand College of Engineering, Sangli, India.

Improved Oversampling Techniques for Handling Imbalanced Big Data Set Classification

Sachin Patil, RIT Rajaramnagar, sachin.patil@ritindia.edu
Shefali Sonavane, WCE Sangli, shefali.sonavane@walchandsangli.ac.in

Abstract — Big Data is a catchphrase of today's research. It depends basically on huge data generated in zettabytes per year. The spotlight is to focus on the analysis of information, by integrating and exploiting it as per the usage. There is a tremendous requirement to study data sets beyond the ability of commonly used software tools and the processing feasibility of the single machine architecture. Training of performance techniques helps to bind the efficient management of Big Data streams. Classification of imbalanced datasets is the point of priority concern in the real world applications. The standard classifiers have produced a notable drawback in the performance of the classification of datasets having imbalanced class distribution. In this paper, a methodology is presented for the handling of single-class/multi-class imbalanced data sets with enhanced over_sampling (O.S.) techniques to improve classification. Various classifiers are further applied on produced balanced data sets to analyze classification. The proposed techniques are applied to imbalanced Big Data (I.B.D.) using mapreduce environment. Experiments are performed on Apache Hadoop using sample datasets. F-measures, ROC area are used to measure the performance of the classification.

Keywords: imbalanced datasets, Big Data, over sampling techniques, data level approach

REFERENCES

[1] B. Chumphol, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safelevel- synthetic minority over-sampling technique for handling the class imbalanced problem," AKDD Springer Berlin Heidelberg, pp. 475-482, 2009.
[2] P. Byoung-Jun, S. Oh, and W. Pedrycz, "The design of polynomial function-based neural network predictors for detection of software defects," Elsevier: Journal of Information Sciences, pp. 40-57, 2013.
[3] N. Chawla, L. Aleksandar, L. Hall, and K. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," PKDD Springer Berlin Heidelberg, pp. 107-119, 2003.
[4] M. A. Nadaf, S. S. Patil, "Performance Evaluation of Categorizing Technical Support Requests Using Advanced K-Means Algorithm," IEEE International Advance Computing Conference (IACC), pp. 409-414, 2015.
[5] R. Sara, V. Lopez, J. Benitez, and F. Herrera, "On the use of MapReduce for imbalanced big data using Random Forest," Elsevier: Journal of Information Sciences, pp. 112-137, 2014.
[6] H. Jiang, Y. Chen, and Z. Qiao, "Scaling up mapreduce-based big data processing on multi-GPU systems," SpingerLink Clust. Comput., vol. 18, no. 1, pp. 369–383, 2015.
[7] N. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321- 357, 2002.
[8] S. Garcia et al., "Evolutionary-based selection of generalized instances for imbalanced classification," Elsevier: Journal of Knowl. Based Syst., pp. 3-12, 2012.
[9] H. Xiang, Y. Yang, and S. Zhao, "Local clustering ensemble learning method based on improved AdaBoost for rare class analysis," Journal of Computational
Information Systems, Vol. 8, no. 4, pp. 1783-1790, 2012.
[10] H. Feng, and L. Hang, "A Novel Boundary Oversampling Algorithm Based on Neighborhood Rough Set Model: NRSBoundary-SMOTE," Hindawi Mathematical Problems in Engineering, 2013.
[11] F. Alberto, M. Jesus, and F. Herrera, "Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning," Springer IPMU, pp. 89–98, 2010.
[12] J. Hanl, Y. Liul, and X. Sunl, "A Scalable Random Forest Algorithm Based on MapReduce," IEEE, pp.849-852, 2013.
[13] D. Chen, Y. Yang, "Attribute reduction for heterogeneous data based on the combination of classical and Fuzzy rough set models," IEEE Trans. Fuzzy Syst., vol. 22, no. 5, pp. 1325–1334, 2014.
[14] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," Elsevier Knowl. Based Syst., vol. 30, pp. 129–135, 2012.
[15] E. Tsang, D. Chen, D. Yeung, and X. Wang, "Attributes reduction using fuzzy rough sets," IEEE Trans. Fuzzy Syst., vol. 16, no. 5, pp. 1130–1141, 2008.
[16] D. Chen, S. Zhao, L. Zhang, Y. Yang, and X. Zhang, "Sample pair selection for attribute reduction with rough set," IEEE Trans. Knowl. Data Engg., vol. 24, no. 11, pp. 2080–2093, 2012.
[17] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," IEE Trans. Fuzzy Syst., vol. 17, no. 4, pp. 824–838, 2009.
[18] D. G. Chen and S. Y. Zhao, "Local reduction of decision system with fuzzy rough sets," Elsevier Fuzzy Sets Syst., vol. 161, no. 13, pp. 1871–1883, 2010.
[19] D. Chen, L. Zhang, S. Zhao, Q. Hu, and P. Zhu, "A novel algorithm for finding reducts with fuzzy rough sets," IEEE Trans. Fuzzy Syst., vol. 20, no. 2, pp. 385–389, 2012.
[20] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," Elsevier Information Sci., vol. 178, pp. 3577–3594, 2008.

[21] W. A. Rivera, O. Asparouhov, "Safe Level OUPS for Improving Target Concept Learning in Imbalanced Data Sets," Proceedings of the IEEE Southeast Con., pp. 1-8, 2015.
[22] S. Yen and Y. Lee, "Under-Sampling Approaches for Improving Prediction of the Minority Class in an Imbalanced Dataset," ICIC 2006, LNCIS 344, pp. 731 – 740, 2006.
[23] M. Bach, A. Werner, J. Żywiec, W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," Elsevier Information Sciences(In Press, Corrected Proof), 2016.

ABOUT AUTHORS

Sachin Patil was born in Mumbai, India, in 1981. He received the B.E. degree in computer science and engineering and M. Tech. in computer science and technology from the Shivaji University, Kolhapur in 2003 and 2011 respectively. He is pursing Ph.D. degree in computer science and engineering under A.I.C.T.E. Q.I.P. scheme at Walchand College of Engineering under Shivaji University, Kolhapur.
Since 2010, he has been an Assistant Professor with the Computer Science and Engineering Department, Rajarambapu Institute of Technology, Rajaramnagar, MH – India. He has worked as head of computer Science and Engineering department at Rajarambapu Institute of Technology, Rajaramnagar, MH – India. He is the author of a book chapter at Springer-Verlag and has more than 15 research papers. His research interests include Database engineering and Big Data analytics. He has received a "Distinguished Facilitator" award at Inspire faculty contest organized by Infosys, Pune.

Dr. Shefali Sonavane is working as Associate professor and heading Department of IT, WCE Sangli. She received research funds from DST and AICTE for various projects supporting work in the area of Computer Vision and Information Security. She has few IPR credentials at her account and extending her research interest further in the fields of Machine Learning and Big Data.

Weblog Analysis Using Hadoop

Gauri D. Kalyankar
gaurikalyankar93@gmail.com
Shivananda R Poojara
shivananda.poojara@ritindia.edul
N V Dharwadkar
nagaraj.dharwadkar@ritindia.edul

Dept. of Computer Science and Engineering
Rajarambapu Institute of Technology, Sakhrale, Sangli Dist

Abstract— In today's world of Internet, websites are convenient source of information. Millions of people are accessing these websites and navigating number of pages by liking, sharing, clicking, and tweeting per day. Hence log files of these websites are increasing in size per day. These log files are significant source of information to recognize the behavior of the user. In current internet world, analysis of log file is becoming a needful task to examine the customer's behavior with the purpose of improvements in advertising and sales. As the web log file size is huge parallel processing and reliable data storage system for processing these files is required. Hadoop is an open source framework which provides parallel and reliable storage and processing of such large datasets. In this work weblog analysis system is implemented by using Hadoop which is capable to process large sizes log files and helps to analyze the output results to study user's behavior.

I. INTRODUCTION

As per the needs of the today's world, everything is going online. Each field is having their own way to put their applications and business online on the internet. Seating in the house anyone is able to do shopping, banking related work; Also able to get weather information, and numerous other services. And in such a competitive environment, different service providers are anxious to know about if they are offering the best services of the market, if people are buying their product, are obtaining interesting and easy to use their application, or in the banking sector they need to know that how many customers are waiting for their bank scheme. Similarly, they also need to know about problems that have been verified, how to solve, how to make web sites or interesting web applications, which products people are not buying, and if so how to improve advertising strategies to attract customers, what should be the future marketing plans. Log files are useful to get the answers of all these questions [1][2][8].

In today's world of Internet, websites are convenient source of information. Millions of people are accessing these websites and navigating number of pages by liking, sharing, clicking, and tweeting per day. Hence log files of these websites are increasing in size per day. To discover the behavior of the user ,these large log files plays vital role. Objective is to study Hadoop Framework in detail and implement Weblog Analysis System using Hadoop.

The data generated in the form of web logs is measures in terms of giga bytes to tera bytes in large there tier web applications and usage of the application is dynamical changeable over the time. Hence its need of an hour to process the logs files for real time analysis of user activities.

Web log consists of data gathered is too huge, dynamic and different variety. Hence it can be considered under problem scenario of big data analytics .Big data analytics is the process which examines large datasets which contains variety of types of data. Big data Analytics is used to uncover hidden information, hidden patterns and to discover knowledge from the large volume data. In th experiment we are using Hadoop framework for data processing and analysis of the web logs.

Overall structure of the paper is as follows Section II briefly describes about prior work. Section III describes the fundamentals of Hadoop eco system and web logs which motivated to conduct the experiment. Section IV indicates the methodology adopted for analysis. Section V describes about results and discussion. Lastly the Section VI concludes the overall research work.

II. PRIOR WORK

Web log analysis is an experimental and distinctive field which is constantly formed and modified by the convergence of many different emerging web technologies. Because of its interdisciplinary nature, the diversity of

the problems it faces. The variety and the number of web applications is the subject of many distinctive and different research methodologies.

Weblog analysis is used to find the statistics from web sites. Weblog analysis can be done using different techniques and statistical tools [6][7].

Weblog analysis system based on Hadoop and its ecosystem was developed, in which authors used MapReduce framework with Pig programming for the efficient results and better performance. Pig programming is used for the effective analysis of log files. The proposed system is equipped with two functions, one for the application server traffic analysis,

and a performance statistics program. The proposed system provides fault tolerance and performance tuning. The main objective of the proposed system is the system helps to quickly capture and analyze the data hidden in enormous potential value and as an important basis for business decisions [1].

The analysis system based on Weka was proposed for the analysis of weblogs .A Weka is open source Java based platform for data mining, which contains different machine learning algorithms which can be used to extract data, including pre-processing of data, classification, clustering, association of mining rules and visual interactive page. The developed system process web logs on Weka platform. Weka processes and analyzes the data, mines uncommon information and finds trends of different users which visit the website and by this it helps to provide support for the decision making and in making structure of website [6].

A framework was developed to analyze weblog social networks built by a query-dependent exploration scheme and query-dependent ranking algorithm. The Weblog social network analysis uses content analysis and semantic issues, and then associate bloggers who never get his interaction. Interactive information visualization techniques such as fish- eye have been introduced for users to explore different features at various levels of abstraction over the underlying social network. Authors have also introduced the applications developed based on the link and content analysis and visualization techniques, which benefits the activities of national security and even the general public to understand each weblog social network[7].

III. HADOOP AND WEBLOG

A. Weblog

When an internet search or link any web site, the web server records the details of it in the form of Web Access Logs. In the Web Access Log you can see the file types, users log on to the place from which the application and other information such as the type of web browser and visitors of the devices in use are made. An access log contains the list of all the items that users have requested from a website. Now days many people are using the internet for their daily activities, such as social networking, news, education and many others. Therefore, websites are acquiring thousands of visitors every day. As a web server keeps record of all user actions, the file size web server log is growing every day. These log files are semi-structured and hence very difficult to store, process and analyze using traditional database systems or traditional data warehouse in minimum time [1][3][8].

B. Hadoop Framework

Apache Hadoop is an open source framework written in Java that allows you to distribute the processing of large amounts of data across clusters of nodes that use simple programming models. Frame-worked Hadoop application works in an environment that provides distributed storage and computing capabilities of computer clusters. Hadoop is designed in such a way that it scales from single server to thousands of nodes, each of which offers the calculation and local storage[1].

1) Hadoop Common: These are the Java libraries, and utilities used by other Hadoop modules. These libraries offers file system and abstraction level of the operating system and contains the necessary Java files and scripts needed to start Hadoop.
2) Hadoop YARN: It is a reference framework for the planning process and the cluster resource management.
3) Hadoop Distributed File System (HDFS): A distributed file system that provides access to high-throughput of application data. The Hadoop Distributed File System (HDFS), is completely based on the Google File System (GFS) and designed to run on large clusters and also provides a distributed file system that is fault-tolerant. HDFS is based on master- slave architecture where the master works as a single Name Node that manages the file system metadata and one or more slave Data Nodes data. The Name Node controls the mapping of blocks to Data Nodes.

The Data Nodes are responsible for reading and writing operations with the file system. They also concerns about the block creation, deletion of blocks and also replication of them based on instructions given by Name Node. HDFS offers a file system to store and access data, and also list of commands which are available to interact with the file system [11].

4) Hadoop MapReduce: This is YARN-based system used for parallel processing of large volume of data sets.

C. MapReduce Framework:

Hadoop MapReduce is a framework that enables ease of writing software applications that are able to process large volume data in parallel on thousands of machines which are commodity hardware, in a authentic and fault-tolerant way [12].

The MapReduce includes these two different phases that are running Hadoop programs as shown in Figure 1:

1) The Map Phase: This is the initial phase that takes the input data and after processing, converts it into intermediate data which is in the form of key/value pairs.
2) The Reduce Phase: This is the second phase in which the output of a map phase is taken as an input and combines the values associated with the key. The task of reducing is always performed after the operation map.
1) Starting Hadoop: After setting all the parameters of system Hadoop can be started by using bin/start-all.sh command. The jps command is used to start NameNode, SecondaryNameNode, SecondaryNode, and JobTracker.
2) Weblog Analysis and Processing: Initial state of system consists running Hadoop infrastructure. Figure 2 shows overall working of Weblog analysis system on Hadoop infrastructure.

HDFS stores both the input and output of the MapReduce framework. The MapReduce framework is able to take care of planning, monitoring the tasks and also re-running the failed tasks [12].

The MapReduce framework composed with a single master Job Tracker and one TaskTracker slave for cluster-node. The master is responsible for the management of resources, monitoring of consumption and availability of resources, the planning of the tasks of the members work on the slaves, their monitoring and re-executing the failed tasks. The TaskTracker slaves execute the tasks as directed by the JobTracker, and provide information about the progress of tasks completion [12].

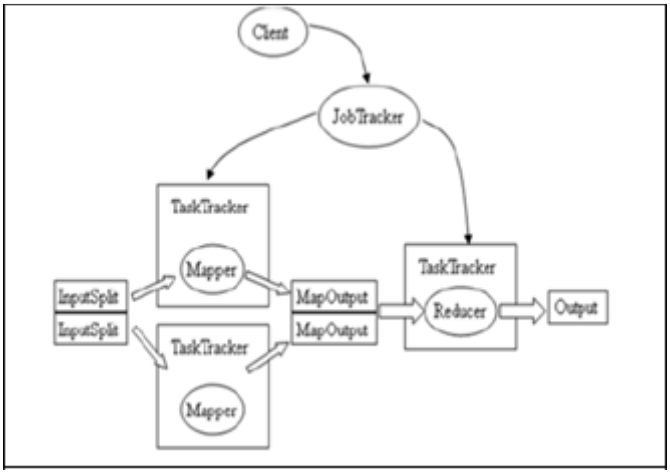


Fig.1. Map Reduce workflow diagram

The Job Tracker is the only point of failure for the Hadoop MapReduce service that means that all running processes are stopped if JobTracker goes down.

IV. METHODOLOGY

The Weblog analysis system is implemented by using Hadoop Framework. Hadoop MapReduce framework is used for processing large size semi structured log files data. The weblog data sets are downloaded from source

sites and converted into text file format for further processing. Python is used for the MapReduce programming. The hardware and software requirements for the implementation of weblog analysis system using Hadoop are

A. Hardware requirements

Personal computer with 1). Processor: i5-4590 @3.30 GHz.2). RAM : 4 GB. 3). HDD: 500 GB.

B. Software requirements

a) VMware Workstation b) Cloudera-quickstart-vm-5.5.0 c) Centos 6.7.

C. Experiment set up and process

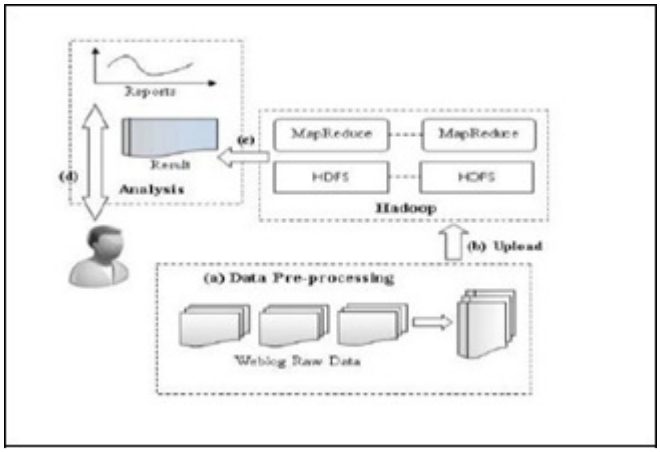


Fig. 2. Web Log Analysis Architecture[1]

In this firstly

[1] The weblog datasets which are in the raw format are converted into text file format as shown in fig 3.

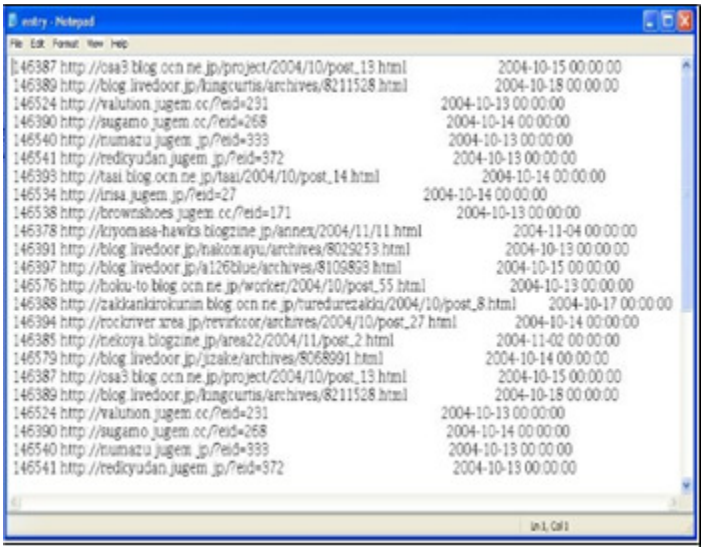


Fig.3. Input Weblog Files in Text File Format

[2] The input file is uploaded on HDFS for processing.

[3] MapReduce Framework is used for processing the input files. In Map phase the input data is converted into inter- mediate data that is in key value pair as shown in fig 4.

The output shows the string as a key is url of a website and the value denotes number of times the particular url is hit by the user in the data. That is the frequency of each url occurred in the particular data.

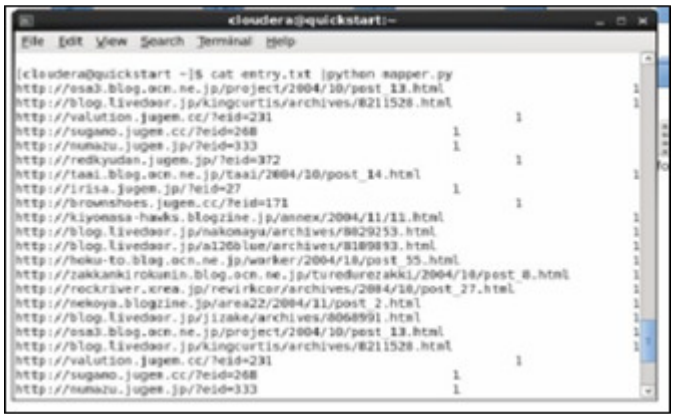


Fig 4. Map phase

[4] In Reduce phase intermediate data is converted into final output by combining all the values associated with a key as shown in fig 5.

Fig 5. Reduce Phase

[5] Output is moved to Linux file system which is stored in specified directory in text file format. After getting result user can get maximum hit count of particular url of a website.

V. RESULTS AND ANALYSIS

By performing weblog analysis results are come in the form of key value pairs as shown in fig 6. The output of the the program is in key value pair as required.

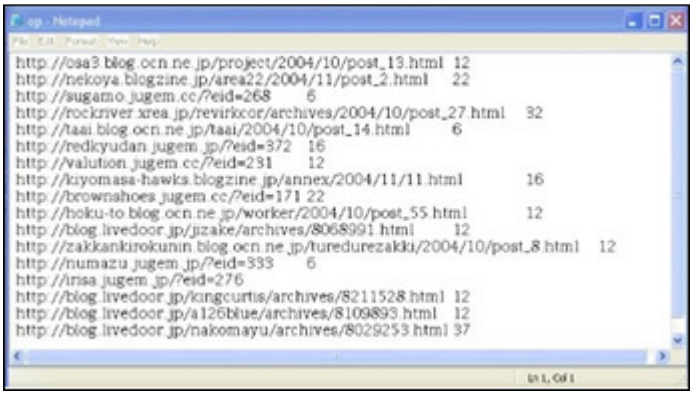


Fig 6. Output Results

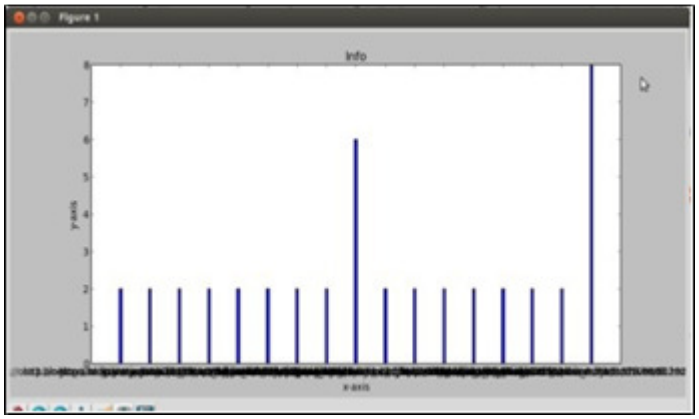


Fig 7. Analysis of Result

From analyzing the results it is possible to conclude about the frequent use of particular url.The graph shown in fig 7 shows the most frequently accessed url. In above graph x-axis denotes link address and y-axis denotes hit count of particular link address. From the graph user can conclude about the frequently accessed url and according to it can make business decisions.

VI. CONCLUSION

In today's internet world, websites are useful source of information. Number of people are accessing to these websites and requesting number of pages, due to which the log files of these websites are getting bigger every day. These log files are convenient source of information to identify the behavior of the user. As developed weblog analysis system is based on Hadoop framework it is capable to process large data sets of weblog. By experiments it is also concluded that by using Hadoop MapReduce framework, current system can perform large size weblog processing and gives efficient results with less efforts and time. As a future work, Hadoop ecosystem like Pig, Flume, Hive for efficient processing of data can be embed in current system to get more effective analysis result and also can use different data visualization techniques to view the results.

REFERENCES

[1] Chen-Hau-Wang,Ching Tsorng Tsai ,Chia-Chan Fan,Shyan-Ming Yuan."A Hadoop based Weblog Analysis System."7th IEEE Conference 2014.
[2] Dhara Kalola, Prof. Uday Bhave."WeblogAnalysis with Map-Reduce and Performance Comparison of Single v/s hadoop Cluster".IJIRCCE.11
[3] Siddharth Adhikari, Devesh Saraf, Mahesh Revanwar, Nikhil Ankam."Analysis of Log Data and Statistics Report Generation Using Hadoop."IJIRCCE 2014.
[4] Sayalee Narkhedeand Tripti Baraskar.HMR "Log Analyzer: Analyze Web Application Logs Over Hadoop MapReduce."IJU July 2013.
[5] Milind Bhandare, Prof. Kuntal Barua, Vikas Nagare, Dynaneshwar Ekhande, Rahul Pawar,"Generic Log Analyzer Using Hadoop MapReduce Framework." IJETAE 2013.
[6] Xiu-yu Zhong."The research and application of web log mining based on the platform weka." Procedia Engineering 15 (2011) 4073 – 4078,ScienceDirect.
[7] Christopher C. Yang and Tobun D. Ng.Terrorism and Crime Related Weblog Social Network:Link, Content Analysis and Information Visualization.
[8] Songxiang Cen, Li han, Jian Ma.Ranking "Weblogs by Analyzing Reading and Commenting Activities". 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops
[9] Lei Wang, Fuji Ren, and Duoqian Miao."A Novel Method for Recognizing Emotions of Weblog Sentences." Proceedings of the 2013 IEEE/SICE International Symposium on System Integration, Kobe International Conference Center, Kobe, Japan, December 15-17
[10] Jianli Duan, Shuxia Liu. "Research on web log mining analysis". IMSNA 2012.
[11] D.Borathkar."The Hadoop Destributed File System:Architecture and Design."The Apache Foundation 2007.
[12] Jeffrey Dean and Sanjay Ghemawat., (2004) "MapReduce: Simplified Data Processing on Large Clusters", Google Research Publication.

SCSP: Skyline Computation Strategic Planner for Distributed Environment

Abstract - A skyline query helps to identify the best objects in a multi-attribute dataset. A typical skyline provider experiences frequent and near to frequent queries on a popular dataset. When the skyline query attributes are more, dataset is very large and frequent updates are common on a dataset, response time can be remarkably improved if the execution of the sub sequent queries is planned with the help of previously available results of the skyline queries.

Through this paper, we present a novel model namely SCSP, a skyline computation strategic planner which is suitable for computing the skylines of queries that are raised against a distributed dataset. The model presents the novel strategies which help to optimize the response time of skyline queries. We then present a novel algorithm based on the SCSP model, the DiSkyline algorithm to optimize the response time of the skylines queries in a distributed environment.

Keywords— Skyline queries; Query profiler; Skyline computation strategic planner

REFERENCES.

[1] R D. Kulkarni, B. F. Momin, "Skyline computation for frequent queries in update intensive environment", In Elsevier journal of King Saud University, Computer and Information Sciences, Volume 28, issue 4, pp. 447-456, 2016.

ABOUT AUTHORS

R. D. Kulkarni completed her M. E. in Computer Science and Engineering from Shivaji University, Kolhapur in 2006 and completed her B. E. in Computer Engineering from University of Pune in 1999. Currently she is a Ph. D. student at Department of Computer Science and Engineering at Shivaji University, Kolhapur. Her area of research is skyline computing.

Dr. B. F. Momin received his Ph.D. from Jadavpur Univerity, Kolkata in 2008 and received his M.E. and B.E. degrees in Computer Science and Engineering from Shijavi University, Kolhapur in 2001 and 1990 respectively. Currently he is Associate Professor at department of degrees in Computer Science and Engineering, Walchande College of Engineering, Sangli. His areas of expertise are database and data mining.

NPL: A tool for managing processor affinity in network

K B MANWADE
Department of Computer Science and Engineering,
Walchand College of Engineering,
Sangli, Maharashtra, India

D B KULKARNI
Department of Information Technology,
Walchand College of Engineering,
Sangli, Maharashtra, India

Abstract—Message Passing Interface (MPI) is a dominant parallel programming paradigm, used in high performance computing (HPC) applications. It is suitable for distributed memory systems, so MPI processes can be placed on HPC systems which contain nodes in the network. Formerly, placing of MPI processes on nodes in network was easy because nodes used to have a single processor. MPI process placement in HPC system has become difficult because of development in multi-core and many-core systems. Such systems contains multiple cores and different levels of memory hierarchy. On such complex systems, topology aware process placement is possible when the architectural and network details of the systems are known. In literature, various tools are proposed to explore the architectural details, but to explore network details no effective tool is available. In this paper, we have proposed a Network Processor Locality (NPL) tool which explores the topology of the system based on parameters like distance between cores, memory hierarchy between cores, network bandwidth, hop count, propagation delay, packet loss ratio, network contention etc.

Keywords—Topology aware MPI process placement, High performance computing, System topology

I. Introduction

Message Passing Interface (MPI) is a dominant programming paradigm used in HPC cluster. While executing MPI applications in cluster, multiple processes are created and scheduled for execution in a cluster. The evolution in multi core and many core technologies [1] made the hardware system more complex. Now a days manufacturing of computer system with hundreds or even thousands of cores is possible. 'Sunway TaihuLight' [2] is the highest ranked HPC system developed at National Supercomputing Center China and it has 10,649,600 cores. On such complex systems placing of MPI processes is difficult. To place processes easily and to get benefits of huge computational power of these systems, the architectural details like number of cores, memory hierarchy etc needs to be considered. Also several modern interconnect technologies like Myrinet, Infiniband, Ethernet-based technologies [3] are available which provide bandwidth in Gigabytes. To take advantage of such huge available bandwidth, processes should be placed by considering system topology. Topology aware MPI process placement [4] involves three steps: (i) finding application topology (ii) finding system topology and (iii) topology aware process placement algorithm. In this paper, the emphasis is given on finding system topology. To construct HPC cluster topology various (compute) node parameters as well as network parameters should be used. In literature various tools are available for finding topology of system but in those only compute node parameters are used and network parameters are not used. We have developed a tool for finding topology of HPC cluster by considering both compute node parameters and network parameters.

The rest of this paper is organized as follows; Section 2 focuses on motivations behind the development of this tool. In section 3 different terminologies related to HPC system topology are elaborated. The implementation details of proposed tool are given in section 4. In section 5, related work is given. In section 6, results are elaborated. Finally in section 7 the conclusion of our work and future plans are given.

II. Motivation

A. Evolution in hardware architecture

There is a huge advancement in architecture of computer systems as illustrated in [5]. The architectural development is mainly in following areas.

- Evolution in processor technology from uniprocessor to dual core processor to multicore processor.
- Hierarchical placement of cores on different socket/s.
- Provision of different level of cache memory.

This is an era of exascale computing where machines with thousands of cores and different levels of memory hierarchies are available. In Table I, list of top five [6] parallel machines is shown. To extract benefits of such huge computing power, processes should be placed by considering architectural details of the node.

TABLE I
Top five multi core servers

RANK	SYSTEM	CORES
1	SUNWAY TAIHULIGHT	10,649,600
2	TIANHE-2	3,120,000
3	TITAN - CRAY XK7	560,640
4	SEQUOIA - BLUEGENE/Q	1,572,864
5	K COMPUTER	705,024

B. Evolution in interconnect technology

Modern interconnect technologies provide bandwidth in gigabytes [6] [7]. Table II shows the bandwidth offered by

various interconnect technologies. MPI applications are not benefited from these technologies, as network parameters are not used while placing processes.

TABLE II
MODERN INTERCONNECT TECHNOLOGIES AND THEIR CORRESPONDING BANDWIDTH

INTERCONNECT TECHNOLOGY	MAX BANDWIDTH (MB/S)
GigE	112
10GigE	875
DDR InfiniBand	1482
QDR InfiniBand	3230

Various network topologies [7] with different network diameter & hop count are available. To improve performance of applications executing in such networks, topology should be considered. Thus, large scale HPC systems are complex as they contain high configuration servers and high speed interconnection networks.

C. Evolution in MPI process placement

The communication latency reduction is one of the technique to improve the performance of MPI applications. Topology aware process placement technique is used for this purpose. Figure 1 shows that, topology aware process placement improves application performance significantly than default process placement. In [8], [9] [10], effects of topology aware process placement on performance of MPI applications are demonstrated. To achieve performance improvement through topology aware process placement, knowing the topology of underlying system is essential.

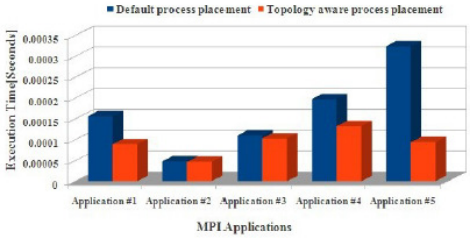


Fig. 1. Performance improvement using topology aware process placement

III. TOPOLOGY OF HPC SYSTEM

A. Parameters of HPC system

HPC system contains compute nodes connected in the network. To construct its topology, compute node parameters as well as the network parameters should be considered. Each parameter has an influence on performance of application executing on the system. Based on the influence of these parameters on application performance, different weights are assigned as shown in Table III. From Table III, first two parameters are of compute node parameters and remaining are network parameters.

TABLE III
PARAMETERS FOR SYSTEM TOPOLOGY AND THEIR CORRESPONDING WEIGHTS

PARAMETERS	WEIGHT
DISTANCE BETWEEN CORES	6
MEMORY HIERARCHY BETWEEN CORES	5
BANDWIDTH BETWEEN NODES	4
HOP COUNT BETWEEN NODES	3
PROPAGATION DELAY BETWEEN NODES	2
PACKET LOSS RATIO BETWEEN NODES	1

The significance of each parameter in deciding the affinity between cores is given below.

1) DISTANCE BETWEEN CORES: The distance parameter indicates the physical distance between cores. If cores are on same socket then distance is 1, if cores are on different socket but on same node then distance is 2 and distance is 3, if cores are on different nodes.
2) MEMORY HIERARCHY BETWEEN CORES: If cores have L1 level cache between them then affinity is 3. If cores have L2 level cache between them then affinity is 2, for L3 it is 1 and for cores on different nodes it is 0.
3) BANDWIDTH BETWEEN NODES: This parameter indicates the available bandwidth between nodes.
4) HOP COUNT BETWEEN NODES: This parameter indicates, number of hops travelled by the message from one node to another node. If nodes are connected under same switch or router then hop count is 1. It is equal to number of switching devices travelled by the message.
5) PROPAGATION DELAY BETWEEN NODES: This parameter indicates time taken by the message to travel from one node to another. The affinity between cores is less when propagation delay is more.
6) PACKET LOSS RATIO BETWEEN NODES: The packet loss cause resend of message, therefore more time is required for transmission of message between nodes. The less packet loss ratio indicates more affinity between nodes.

B. Affinity between cores

While defining affinity between cores in multicore and network based system various parameters and their impact on processor affinity should be considered.The affinity between cores can be expressed as follows.

$$(c_i, c_j) \quad Clust_A \quad \text{where} \quad i = j = k \quad \text{if} \tag{1}$$

$$\text{(i)} \quad dist(c_i, c_j) < dist\{(c_i, c_k), (c_j, c_k)\} \tag{2}$$

$$\text{(ii)} \quad mem - hier.(c_i, c_j) < mem - hier.\{(c_i, c_k), (c_j, c_k)\} \tag{3}$$

$$\text{(iii)} \quad bandwidth(c_i, c_j) > bandwidth\{(c_i, c_k), (c_j, c_k)\} \tag{4}$$

$$\text{(iv)} \quad hop - count(c_i, c_j) < hop - count\{(c_i, c_k), (c_j, c_k)\} \tag{5}$$

$$\text{(v)} \quad prop. - delay(c_i, c_j) < prop. - delay\{(c_i, c_k), (c_j, c_k)\} \tag{6}$$

(vi) $$\text{packet} - \text{loss}(c_i, c_j) < \text{packet} - \text{loss}\{(c_i, c_k), (c_j, c_k)\} \quad (7)$$

Let us assume that, c_i, c_j and c_l are the cores under consideration. Core c_i and c_j can be grouped in same cluster $clust_k$ if they have less distance, memory hierarchy, propagation delay, hop count, packet loss ratio and more bandwidth between them.

C. Topology of HPC system

To represent topology of HPC system we have used weighted graph. Let G(V,E,W) be the weighted graph where V is a set of compute nodes, E be the set of edges between compute nodes and W be the set of weights on edges. To store values of each parameter, matrices are used. Let S be the system topology matrix, its values are calculated as,

$$\begin{aligned} S[i][j] = A[i][j] \times 6 + B[i][j] \times 5 + C[i][j] \times 4 \\ + D[i][j] \times 3 + E[i][j] \times 2 + F[i][j] \times 1 \end{aligned} \quad (8)$$

Where, A, B, C, D, E, F are the matrices of system parameter.

The dimension of node topology matrix depends on number of cores in the node and dimension of network topology matrix depends on number of nodes in the system. The dimension of system topology matrix is dependent on both number of nodes in the system and number of cores in node. So, the dimension (Dim) for system topology matrix is,

$$\begin{aligned} Dim = MAX\,dimension\{A, B, C, D, E, F\} \times \\ MAX\,processors\{A, B, C, D, E, F\} \end{aligned} \quad (9)$$

D. Data structure for topology

Topology aware process to core mapping is NP-complete problem as shown in [11]. The mapping time depends on data structure used for representation of the system topology and application topology. Therefore we have used a data structure called as dendrogram [12] to represent topology of system. The key benefits of this data structure are,

- It explores the hierarchical nature of the system
- It forms clusters of cores based on various parameters of node as well as network and hence finds the nearest cores.
- Its tree like structure reduces the mapping time
- It is compatible with existing work as it works on data in matrix format

The sample system topology is shown in figure 2 and its corresponding topology in dendrogram format is shown in figure 3.

As shown in figure 2, the system contains four nodes 'Node1', 'Node2', 'Node3' and 'Node4'. Each node has two sockets and on each socket four cores are situated. The 'dendrogram' groups cores socket wise (G1 to G8), node wise (G9 to G12) and system wise (G13) as shown in figure 3.

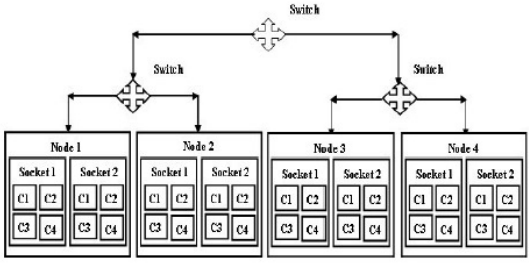


Fig. 2. Sample system topology

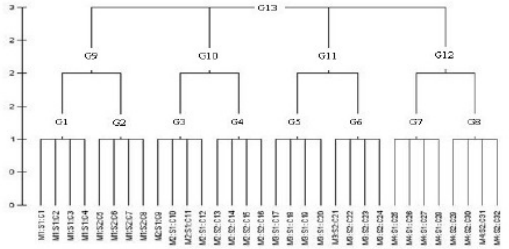


Fig. 3. Sample system topology in Dendrogram format

IV. Network Processor Locality (NPL) tool

A. Architecture of NPL

The architecture of NPL consists of following four layers as shown in figure 4.

1) Layer1: It contains nodes from HPC system and network components. The HPC nodes are servers having complex structure of multiple cores and memory hierarchy. The network components are different interconnect technologies and networking devices such as switch, router, hub etc.
2) Layer2: It contains tools which are available to explore the topological details of nodes and network. The hwloc tool is used to extract parameters related to the architectural details of a node such as number of cores, physical locality of core, memory hierarchy, size of cache etc. Linux utilities like ping, traceroute, lscpu, lspwd, /etc/proc, etc/sys are used to extract network parameters like bandwidth, propagation delay, packet loss ratio and hop count.
3) Layer3: It contains our tool which merges the topology information of nodes, network and produces topological information for entire HPC system. This topological information is stored in matrix format. While producing the final topology matrix weighted sum of all parameters is taken. Weights are assigned to each parameter based on its impact in deciding affinity between the cores.
4) Layer4: It contains a hierarchical clustering tool called as 'multidendrogram' [12] which creates cluster of nodes.

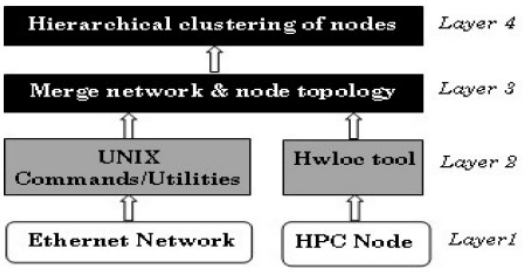


Fig. 4. Architecture of NPL

B. Implementation of NPL

To get information for system parameters, we have used basic utilities from 'linux' and 'hwloc' [13]. The summary of parameters and corresponding utilities is shown in Table IV.

TABLE IV
SYSTEM TOPOLOGY PARAMETERS AND THEIR CORRESPONDING TOOLS

PARAMETER	UTILITIES USED
BANDWIDTH BETWEEN NODES	ping, ipconfig
PROPAGATION DELAY BETWEEN NODES	ping
PACKET LOSS RATIO BETWEEN NODES	ping
DISTANCE BETWEEN CORES	lstopo, lscpu
MEMORY HIERARCHY BETWEEN CORES	lstopo, lscpu

The obtained values of parameters are stored in matrices. The system topology matrix is constructed by taking weighted sum of values from node and network parameter matrices.

Algorithm 1: MergeTopology

Input:
A //Distance matrix(Node parameter)
B //Mem. Hierarchy matrix (Node parameter)
C //Bandwidth matrix (Network parameter)
D //Hop_count matrix (Network parameter)
E //Propagation_delay matrix (Network parameter)
F //Packet_loss matrix (Network parameter)
Output:
S //System topology matrix
Procedure:
Begin
1. A'=ExpandNodeMatrix (A)
2. B'=ExpandNodeMatrix (B)
3. C'=ExpandNetworkMatrix (C)
4. D'=ExpandNetworkMatrix (D)
5. E'=ExpandNetworkMatrix (E)
6. F'=ExpandNetworkMatrix (F)
7. For i=0; i < Dim(S); i++ //Dim(S) is a dimension of system topology matrix and is given in equation (2)
8. For j=0; j < Dim(S); j++
9. S[i][j] = A' [i][j] x 6 + B'[i][j] x 5 + C' [i][j] x 4 + D' [i][j] x 3 + E' [i][j] x 2 + F'[i][j] x 1
//As shown in equation (1)
End

Procedure: ExpandNodeMatrix (X)

X : Node parameter matrix
Begin
1. For i=0 ; i < Dim(X) ; i++
2. For j=0 ; j < Dim(X) ; j++
3. k=j
4. X'[i][j] = X[i][j]
5. X'[i][j+k] = X[i][j]
6. X'[i+k][j] = X[i][j]
7. X'[i+k][j+k] = X[i][j]
8. Return X'
End

Procedure: ExpandNetworkMatrix (Y, d)

Y : Network parameter matrix
d[]: Array of number of cores per node
Begin
1. cores = 0
2. For i=0 ; i < Dim(Y) ; i++
3. For j=0 ; j < d[i] ; j++
4. For k=0 ; k < d[i] ;k++
5. Y'[j+cores] [k+cores] = Y[j][k]
6. cores = cores + d[i]
7. Return Y'
End

To construct system topology matrix, we have used node parameters as well as network parameters. The node parameter matrices are focusing on details of node and network parameter matrices are focusing on details of network. To construct the system topology matrix, expansion and merging of these matrices is required. To expand network parameter matrix we have developed 'ExpandNetworkMatrix' algorithm. The process of expansion is demonstrated in figure 5, where 4x4 network parameter matrix is expanded to 8x8 matrix. Similarly to expand node parameter matrices, we have developed 'ExpandNodeMatrix' algorithm. The process of expansion is shown in figure 6. After expansion of these matrices, the system topology matrix is constructed using 'MergeTopology' procedure from algorithm 1. As the time complexity of NPL tool is $O(N^2)$, it scalable with respect to number of nodes in the system.

V. RELATED WORK

As shown in [14], various tools are available in Linux for finding system topology. These tools provide information about CPU, cache, PCI, data bus, CPU clock etc. To find topology of system, intergeneration of such information from all nodes is required. These tools are useful for node parameters only.

Node	N1	N2	N3	N4
N1	0	1	1	1
N2	1	0	1	1
N3	1	1	0	1
N4	1	1	1	0

Core	N1:C1	N1:C2	N2:C1	N2:C2	N3:C1	N3:C2	N4:C1	N4:C2
N1:C1	0	0	1	1	1	1	1	1
N1:C2	0	0	1	1	1	1	1	1
N2:C1	1	1	0	0	1	1	1	1
N2:C2	1	1	0	0	1	1	1	1
N3:C1	1	1	1	1	0	0	1	1
N3:C2	1	1	1	1	0	0	1	1
N4:C1	1	1	1	1	1	1	0	0
N4:C2	1	1	1	1	1	1	0	0

Fig. 5. Expansion of network parameter

Core	N1:C1	N1:C2	N2:C1	N2:C2	N3:C1	N3:C2	N4:C1	N4:C2
N1:C1	0	1						
N1:C2	1	0						
N2:C1			0	1				
N2:C2			1	0				
N3:C1					0	1		
N3:C2					1	0		
N4:C1							0	1
N4:C2							1	0

Core	N1:C1	N1:C2	N2:C1	N2:C2	N3:C1	N3:C2	N4:C1	N4:C2
N1:C1	0	1	3	3	3	3	3	3
N1:C2	1	0	3	3	3	3	3	3
N2:C1	3	3	0	1	3	3	3	3
N2:C2	3	3	1	0	3	3	3	3
N3:C1	3	3	3	3	0	1	3	3
N3:C2	3	3	3	3	1	0	3	3
N4:C1	3	3	3	3	3	3	0	1
N4:C2	3	3	3	3	3	3	1	0

Fig. 6. Expansion of node parameter

The 'hwloc' [10] tool is the most suitable tool for NUMA architectures. It is developed by combining features of both libtopology [15] and PLPA [16] tools. It provides utilities for integrating topology of various nodes. In this network parameters are not used.

The LIKWID [17] tool provide hardware performance counter and its function is limited to one node only. It does not provide information required for calculating affinity between cores.

The 'servet' [18] tool benchmarks UMA compute nodes by assessing cache hierarchy, memory access cost, memory bottleneck and communication latency. As only node parameters are considered, this tool is not suitable for network based system.

In [19], authors have used PM2 runtime tool to extract node topology. In this tool only node parameters are used. Network locality ('netloc') [20] tool is an extension of hwloc tool which provides the topology of entire system in user readable format. But in this, only network parameters are not used.

In [5], authors have used LibTopoMap tool to discover the topology of system. In this tool network parameters are considered but node parameters are not used.

As HPC system contains both nodes and network, both node and network parameters should be considered.

VI. Results

A. System under consideration

To test our NPL tool, we have used system having four nodes with dual core processors and are connected with Ethernet network as shown in figure 7. The network topology used is star topology and connection media used is Cat5 cable.

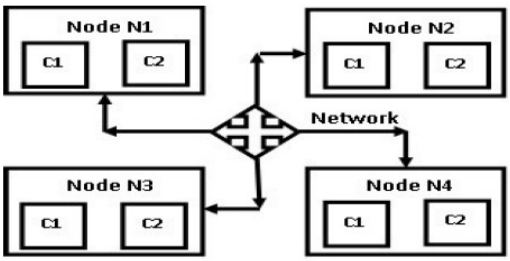


Fig. 7. Sample network based system

B. System topology using NPL

The information from all nodes in the system is collected in individual matrices. The weighted sum (where weightage of each parameter is considered) of all matrices are stored in system topology matrix.

TABLE V
A) Bandwidth matrix (Bandwidth in Mbps) b) Hop count matrix

Node	N1	N2	N3	N4	Node	N1	N2	N3	N4
N1	260.0	11.0	32.85	2.15	N1	0.0	1.0	1.0	1.0
N2	27.0	3285.0	3838.0	2.15	N2	1.0	0.0	0.0	1.0
N3	2.15	27.0	11.0	5750.0	N3	1.0	0.0	0.0	1.0
N4	2.15	27.0	11.0	5750.0	N4	1.0	1.0	1.0	0.0

In Table V(a), available bandwidth between nodes and in Table V(b) hop count between nodes are shown. If two nodes are under same switch then hop count is zero otherwise it is equal to number of switches between them. Similarly matrices for message propagation delay between nodes, packet loss ratio between nodes, distances between cores, memory hierarchy between cores, system topology matrix are stored. The clusters of nearest cores are created by using 'multidendrogram tool. The clusters are shown using hierarchical clustering technique and dendrogram data structure. The system topology is shown in figure 8. Cores in machine N1 & N4, N2 & N3 are closer; hence they are clustered in same cluster.

C. Performance analysis

For the system shown in figure 7, ten clusters [(N 1, N 2), (N 1, N 3), (N 1, N 4), (N 2, N 3), (N 3, N 4), (N 1, N 2, N 3), (N 1, N 2, N 4), (N 1, N 3, N 4), (N 2, N 3, N 4), (N 1, N 2, N 4)] of nodes can be created with permutation combination and by using minimum distance parameter. By using our tool, we have

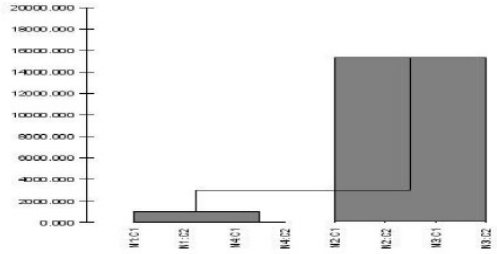


Fig. 8. Dendrogram for system topology

created two clusters as (N 1, N 4) and (N 2, N 3). We have analysed clustering performance of our tool by using minimum distance parameter. As shown in figure 9, our tool creates clusters (N 1, N 4) and (N 2, N 3) with minimum distance which reduces inter node communication latency. Thus it improves performance of application in case of topology aware process placement. Also the accuracy of

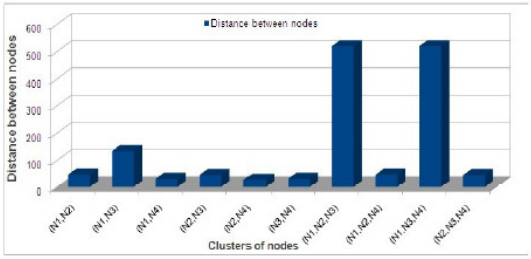


Fig. 9. Performance analysis of NPL tool

NPL tool is compared with existing 'hwloc' and 'netloc' tools by using 'total intra cluster bandwidth' parameter. The results are given in Table VI. NPL creates clusters of processors/cores with maximum intra cluster bandwidth as compared to 'hwloc' and 'netloc' tools.

TABLE VI
COMPARISON BETWEEN NPL, HWLOC AND NETLOC

Tool	Clusters obtained	Total Intra cluster bandwidth (Mbps)
NPL	(N1,N4), (N2,N3)	160487
hwloc	(N1),(N2),(N3),(N4)	106674
netloc	(N1),(N2),(N3),(N4)	106674

VII. CONCLUSION

As the HPC system consists of nodes connected through network, the network parameters plays an important role in deciding the affinity between the cores in entire system. In our tool we have considered network parameters and their corresponding weights in deciding the affinity between cores. We have obtained correct clustering of cores. If only distance between cores, memory hierarchy are considered like 'hwloc' and 'netloc' tool, wrong clustering of cores would be generated. Thus performance of MPI program will be not improved. We have come to the conclusion that for standalone system it is sufficient to use tools like 'hwloc' and 'netloc', but for network based system these tools are not effective. Our tool is useful for both standalone as well as network based system. We are planning to extend our tool for heterogeneous system which contains different types of hardware architectures. We want to extend this tool for hybrid system which contains NUMA as well as GPU architecture. Our tool works for Ethernet network and we are planning to extend it for InfiniBand network as well.

REFERENCES

[1] Vajda, Andrs, Programming Many-Core Chips, Pages 9-43, Springer, 2011.
[2] https://www.top500.org/lists/2016/06/
[3] Marco Vanneschi, Interconnection Technologies for High-Performance Distributed Systems, Course Notes on High Performance Computing, 2014.
[4] Hari Subramoni, Topology-Aware MPI Communication and Scheduling for High Performance Computing Systems, Ph.D. thesis, The Ohio State University, 2013.
[5] Balaji Venu, Multi core processors An overview, Department of Electrical Engineering and Electronics, University of Liverpool,Liverpool,UK, 2010.
[6] HPC Advisory Council, Interconnect Analysis: 10 GbE and InfiniBand in High Performance Computing, HPC Advisory Council, 2009.
[7] Dennis Abts and John Kim, High Performance Datacenter Networks: Architectures, Algorithms, and Opportunities,Lectures on Computer Architecture, Morgan and Claypool Publishers series, 2011.
[8] Albert Solernou, Jeyarajan Thiyagalingam, Mihai C. Duta, Anne E. Trefethen, The Effect of Topology-Aware Process and Thread Placement on Performance and Energy, In Proceedings 28th International Supercomputing Conference, pp 357-371, 2013.
[9] Emmanuel Jeannot and Guillaume Mercier, Near-Optimal Placement of MPI processes on Hierarchical NUMA Architectures, 16th International Euro-Par Conference on Parallel Processing, 2010.
[10] Guillaume Mercier and Jerome Clet-Ortega, Towards an Efficient Process Placement Policy for MPI Applications in Multicore Environments, 16th Euro PVM/MPI conference, 2009.
[11] T. Hoefler and M. Snir, Generic Topology Mapping Strategies for Large-scale Parallel Architectures, In Proceedings of the 2011 ACM International Conference on Supercomputing , pages 75-85, 2011.
[12] Alberto Fernndez and Sergio Gmez, Solving Non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms, Journal of Classification, Springer, pages 43-65, 2008.
[13] Francois Broquedis and et al, hwloc: a Generic Framework for Managing Hardware Affinities in HPC Applications, 18th Euromicro International Conference on Parallel, Distributed and Network- Based Computing, 2010.
[14] Brice Goglin, Managing the Topology of Heterogeneous Cluster Nodes with Hardware Locality (hwloc). International Conference on High Performance Computing and Simulation, Jul 2014.
[15] http://runtime.bordeaux.inria.fr/libtopology/
[16] Portable Linux Processor Affinity (PLPA). http://www.open-mpi.org/projects/plpa.
[17] Treibig J, Hager G, Wellein G, LIKWID: a lightweight performance oriented tool suit for x86 multicore environment, International conference on parallel processing workshop, Washington, USA, 2010.
[18] Gonzalez-Dominguez J et al, Servet: a benchmark suite for auto tuning on multicore cluster, IPDPS, 2010.
[19] Emmanuel Jeannot, Guillaume Mercier, Francois Tessier, Process Placement in Multicore Clusters: Algorithmic Issues and Practical Techniques, Research Report RR-8269, INRIA, pp.32, 2013.
[20] Brice Goglin, Joshua Hursey, Jeffrey M. Squyres, netloc: Towards a Comprehensive View of the HPC System Topology, Fifth International Workshop on Parallel Software Tools and Tool Infrastructures,2014.

Parallel High Performance Content Based Publish-Subscribe System

Mrs. Medha A. Shah, medha.shah@walchandsangli.ac.in
Department of Computer Science and Engineering
Dr. D. B. Kulkarni, dinesh.kulkarni@walchandsangli.ac.in
Department of Information Technology, Walchand Collage of Engineering, Sangli (M.S), India

Abstract- Publish-subscribe model is a loosely-coupled paradigm where applications interact indirectly and asynchronously. This model does not maintain fixed communication channel as messages are routed based on their content. Publish/subscribe model is a natural choice for designing large scale systems, where applications from different domains communicate via middle-ware solutions. The key problems in content based publish/subscribe (pub-sub) system lies in an efficient matching of an event against a large number of subscribers on a single message broker. Existing matching algorithms are inherently sequential and unable to exploit multicore, many-core and accelerator framework readily available in current generation computers. Here we present a novel hybrid model for parallel event processing by combining message passing interface and CUDA (MPI-CUDA), the parallel computing platform and programming model, invented by NVIDIA. We compare our work with CCM (Cuda Content Matching Algorithm), a high performance, and parallel content matching algorithm. The result shows approximately 32% improvement in matching latency. This approach works well where the rate of arrival of an event is significant.

Keywords— Matching Latency, MPI-CUDA, High performance, parallel event processing.

REFERENCES

[1] Alessandro Margara, Gianpaolo Cugola, "High-Performance Publish-Subscribe Matching Using Parallel Hardwarel", IEEE Transactions on Parallel and Distributed Systems Volume: 25, Issue: 1 January 2014.
[2] A. Farroukh, E. Ferzli, N. Tajuddin, and H.-A. Jacobsen, "Parallel Event Processing for Content-Based Publish/Subscribe Systems", (DEBS '09), pp. 8:1-8:4, 2009.
[3] Raphael Barazzutti, Pascal Felber. Streamhub, "A Massively Parallel Architecture for High-Performance Content-Based Publish/Subscribe", DEBS'13, June 29– July 3, 2013, Arlington, Texas, USA.
[4] Medha A. Shah, D.B.Kulkarni, "Storm Pub-Sub: High Performance, Scalable Content Based Event Matching system using Storm", IPDPS'2015
[5] Zhaoran Wang, Xiaotao Chang., "Pub/Sub on Stream: A Multi-Core Based Message Broker with Qos Support", DEBS 2012, July 16–20, 2012, Berlin, Germany July 2012.Forman, G. 2003.
[6] Medha Shah, D.B.Kulkarni, "Enabling Qos Support for Multi-Core Message Broker in Publish/Subscribe System", Advance Computing Conference (IACC), 2014, IEEE International conference.
[7] K.H. Tsoi, I. Papagiannis, M. Migliavacca, W. Luk, and P Pietzuch, "Accelerating Publish/Subscribe Matching on Reconfigurable Supercomputing Platforms", Proc. Many-Core and Reconfigurable Supercomputing Conf., 2010.
[8] A. Carzaniga, D. S. Rosenblum, and A. L. Wolf, "Design and evaluation of a wide-area event notification service", ACM TCS, 2001.
[9] M. K. Aguilera, R. E. Strom, D. C. Sturman, M. Astley, and T. D. Chandra, "Matching events in a content-based subscription system", In PODC, 1999.and evaluation of a wide-area event notification service. ACM TCS, 2001.
[10] H.-A. Jacobsen, A. Cheung, G. Lia, B. Maniymaran, V. Muthusamy, and R. S. Kazemzadeh, "The PADRES publish/subscribe system", Handbook of Research on Adv. Dist. Event-Based Sys. Pub. /Sub. and Message Filtering Tech., 2009.
[11] Peter R. Pietzuch, ―Hermes: "A scalable event-based middleware", Technical Report University of GANBRIDGG Computer Laboratory.
[12] G. Cugola and G. Picco. REDS, "A Reconfigurable Dispatching System", In SEM, pages 9—16, Portland, 2006. ACM Press
[13] T. Yan and H. Garcia-Molina, "Index structures for selective dissemination of information under the Boolean model", ACM TODS'94.
[14] F. Fabret, H.-A. Jacobsen, F. Llirbat, J. Pereira, K. A. Ross, and D. Shasha, "Filtering algorithms and implementation for fast pub/sub systems", SIGMOD'01.
[15] S. Whang, C. Brower, J. Shanmugasundaram, S. Vassilvitskii, E. Veer, R. Yerneni, and H. Garcia-Molina, "Indexing Boolean expressions", In VLDB'09.
[16] M. K. Aguilera, R. E. Strom, D. C. Sturman, M. Astley, and T. D. Chandra, "Matching events in a content-based subscription system", In PODC'99.
[17] A. Campailla, S. Chaki, E. Clarke, S. Jha, and H. Veith, "Efficient filtering in publish-subscribe systems using binary decision diagrams", In ICSE'01
[18] G. Li, S. Hou, and H.-A. Jacobsen, "A unified approach to routing, covering and merging in publish/subscribe systems based on modified binary decision diagrams", In ICDCS'03.
[19] Tania Banerjee Mishra, Sartaj Sahni, "PUBSUB: An Efficient Publish/Subscribe System", IEEE Transactions on Computers (Volume: 64, Issue: 4), 2014.
[20] Mohammad Sadoghi, Hans-Arno Jacobsen, "Analysis and Optimization for Boolean Expression Indexing", ACM Transactions on Database Systems, Vol. 38, No. 2, Article 8, Publication date: June 2013.
[21] M. Sadoghi and H.-A. Jacobsen, "BE-Tree An Index Structure to Efficiently Match Boolean Expressions over High-dimensional Discrete Space", SIGMOD 2011. https://www.linux.com/blog/building-beowulf-cluster just-13-steps.
[22] A. Carzaniga and A.L. Wolf, "Forwarding in a Content-Based Network", Proc. SIGCOMM, pp. 163-174, 2003.

ABOUT AUTHORS

Mrs. Medha A. Shah received M.E (CSE) Degree from Shivaji, University, Kolhapur in year 2008. She is currently Ph.D. student, at Walchand College of Engineering, Sangli. Her area of research is High Performance Computing. At preset she is working as Asst. Professor in CSE Dept. of Walchand College of Engineering, Sangli. Her main research area includes design of parallel algorithms for content based distributed systems. She is author of some papers in the area of parallel and distributed systems, published in international journal and conference proceedings. She has presented her research work in international conferences held at Irvine, USA and Oslo, Norway.

Dr. D. B. Kulkarni received the doctorate degree in Computer Science and Engineering from Shivaji University, Kolhapur in the year 2005. Currently he is a professor in Information Technology Department of Walchand college of Engineering, where he teaches many courses in the area of Computer Science. During his professional life, he has been involved in several R&D projects funded by AICTE, DRDO and UGC. Recently he received Early Adopter award of National Science Foundation's (NSF) of Technical Committee on Parallel Processing (TCPP) of US$1500 for framing and adapting curriculum on “Parallel Computing” in UG and PG curriculum in the institute. He was Visiting Professor in institute of Computer Technology (ICT) of Vienna University of Technology, Austria, in 2010. He is coauthor of tens of scientific papers published in international journals and conference proceedings. His research interests include the area of High Performance computing and Computer Network.

Statistical Machine Translation: Foundations, Challenges and Recent Advances

Arun R Babhulgaonkar, arbabhulgaonkar@dbatu.ac.in
Shefali P. Sonavane, shefali.sonavane@gmail.com
Department of Computer Sci. & Engineering
Walchand College of Engineering, Sangli, Maharashtra, India

***Abstract*—**In general, statistical machine translation (SMT) treats the translation of natural language as a machine learning problem. By examining test samples of human produced conversion, SMT algorithms automatically learn the mechanism. It is observed that, SMT has made tremendous strides in less than two decades and new ideas are frequently introduced. This survey paper presents an overview of the state of the art in the area. The paper describes current research for two main problems namely alignment modeling problem and language modeling problem of phrase based SMT. Along the way, the paper presents taxonomy of few approaches within concerned areas and also enlists some other difficulties in SMT techniques.

Keywords —Machine translation (MT), phrase based machine translation, word alignment, language modeling.

BIBLIOGRAPHY

[1] P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, R.L. Mercer: The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics, Vol. 19, No. 2, pp. 263–311, June 1993.
[2] K.A. Papineni, S. Roukos, R.T. Ward: Maximum Likelihood and Discriminative Training of Direct Translation Models. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, pp. 189–192, Seattle, WA, May 1998.
[3] Y. Al-Onaizan, J. Curin, M. Jahr, K. Knight, J.D. Lafferty, I.D. Melamed, D. Purdy, F.J. Och, N.A. Smith, D. Yarowsky: Statistical Machine Translation, Final Report, JHU Workshop, 1999. http://www.clsp.jhu.edu/ws99/projects/mt/final report/mt-final-report.ps.
[4] F.J. Och, R. Zens, H. Ney: Efficient Search for Interactive Statistical Machine Translation. Proc. 10th Conf. of the Europ. Chapter of the Assoc. for Computational Linguistics (EACL), pp. 387–393, Budapest, Hungary, April 2003.6
[5] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga: A DP-based Search Using Monotone Alignments in Statistical Translation. Proc. 35th Annual Meeting of the Assoc. for Computational Linguistics (ACL), pp. 289–296, Madrid, Spain, July 1997.
[6] F.J. Och, H. Ney: The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics, Vol. 30, No. 4, pp. 417–449, December 2004.
[7] J. Tom´as, F. Casacuberta: Monotone Statistical Translation using Word Groups. Proc. Machine Translation Summit VIII, pp. 357–361, Santiago de Compostela, September 2001.
[8] D. Marcu, W. Wong: A Phrase-Based, Joint Probability Model for Statistical Machine Translation. Proc. Conf. on Empirical Methods for Natural Language Processing (EMNLP), pp. 133–139, Philadelphia, PA, July 2002
[9] Birch, C. Callison-Burch, M. Osborne, P. Koehn: Constraining the Phrase-Based, Joint Probability Statistical Translation Model. Proc. Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL): Workshop on Statistical Machine Translation, pp. 154–157, New York City, NY, June 2006.
[10] C. Tillmann, F. Xia: A Phrase-based Unigram Model for Statistical Machine Translation. Proc. Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), Companion Volume: Short Papers, pp. 106–108, Edmonton, Canada, May/June 2003.
[11] P. Koehn, F.J. Och, D. Marcu: Statistical Phrase-Based Translation. Proc. Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics Annual Meeting (HLT-NAACL), pp. 127–133, Edmonton, Canada, May/June 2003

ABOUT AUTHORS

Babhulgaonkar A. R. is a QIP Ph.D. scholar in Computer Science & Engineering Department of Walchand College of Engineering, Sangli. He is Assistant Professor in Department of Information Technology of Dr. Babasaheb Ambedkar Technological University, Lonere. His research interests include machine translation and machine learning.

Dr. Shefali Sonavane is working as Associate professor and heading Department of IT, WCE Sangli. She received research funds from DST and AICTE for various projects supporting work in the area of Computer Vision and Information Security. She has few IPR credentials at her account and extending her research interest further in the fields of Machine Learning and Big Data.

Underwater Image Enhancement and Color Correction with Wavelet Fusion

Sonali Sankpal and Shraddha Deshpande

Abstract - As water is 800 times denser than air, degradation of underwater images is different from normal images. In underwater imaging light interact with water in two ways, scattering and absorption. The scattering and absorption of light together is called attenuation of light. This attenuation is wavelength dependent. Because of scattering and absorption, underwater images are affected by back scattering, forward scattering and absorption of light in water. As a result of which the images are degraded, and also dominated by green or blue color. Water is not only source of degradation, but underwater images are also affected by suspended particles and dissolved compounds in water. The degradation also depends upon depth of water, day time, geographical location, source of light and physical properties of water etc. Many underwater applications need clear underwater images. Clear underwater images can be achieved by image enhancement. Underwater image enhancement techniques includes contrast enhancement, blur removal, and color correction techniques. This paper proposed the new method for contrast enhancement and color correction with wavelet fusion. The image quality metrics used in this paper are based on primary and fundamental vision parameters perceived by human vision system. The method is compared quantitatively with some state of art methods.

Keywords: underwater image enhancement, color correction, haze removal, contrast enhancement, wavelet fusion.

AUTHORS PROFILE

Mrs. Sonali S. Sankpal received Masters Degree in Electonics and Telecommunication Engineering from Shivaji University, Kolhapur, Maharashtra, India. She is working as Assistant Professor in PVPIT, Budhgaon, Sangli, Maharashtra, India. She is research scholar of Walchand College of Engineering, under in Shivaji University, Kolhapur, Maharashtra, India. Her area of research is image processing.

Dr. Mrs. Shraddha S. Deshpande received Ph. D. Degree from IIT Mumbai. She is Professor in the Department of Electronics Engineering at Walchand College of Engineering Sangli, Maharashtra, India. Her research interest are Optimization Techniques, Model Predictive Control, Instrumentation, is image processing. She is also recognized Ph. D. guide in Shivaji University, Kolhapur, Maharashtra, India.

Fuzzy Extreme Learning Machine with Attribute Selection

Archana P. Kale, Department of Computer Science and Engineering
Shefali P. Sonavane, Department of Information Technology
Walchand College of Engineering, Sangli

Abstract **-** Classification is one of the major thrusts in the field of Machine Learning and Pattern Recognition. In this research work, the issues of learning speed and generalization performance in classification; attribute subset selection as well as fast learning approaches are discussed. Extreme Learning Machine is one of fastest learning algorithms because the weights between input layers to the hidden layer are assigned randomly and weights between hidden layers to the output layer are determined analytically. Extreme Learning Machine has various advantages like speed, reliability and generalization over traditional neural network classifiers.

Fuzzy Extreme Learning Machine with various attribute subset selection is the main idea of this proposed work. Attribute subset selection methods such as Filter, Wrapper, Hybrid, Embedded and Ensemble are used for evaluation. The main objective of subset selection is to provide faster and more cost-effective model. This also gains, a deeper insight into the underlying processes that generate the data. Clinical datasets such as Statlog Heart Disease and Pima Indian Diabetes dataset from University of California Irvine machine repository are used for experimentation. Simulation results are calculated by using Waikato Environment for Knowledge Analysis (WEKA) and MATLAB 14. In general, the simulation results on the benchmark dataset have shown that the proposed FELM classifier with attribute subset selection is immensely efficient with good performance in undertaking the pattern classification task.

REFERENCES

[1] Gulsah Karakaya, Stefano Galelli, Selin Damla Ahipasaoglu and Riccardo Taormina, "Identifying (Quasi) Equally Informative Subsets in Feature Selection Problems for Classification: A Max-Relevance Min-Redundancy Approach", IEEE Transaction on Cybernetics, vol. 46, no. 6, pp.1424-1437, June 2016.
[2] Bharathi P. T. and Dr. P. Subashini, "Optimal Feature Subset Selection Using Differential Evolution with Sequential Extreme Learning Machine for River ice Images", IEEE Conference on Cybernetics, vol. 2, no.1, pp. 1-6, 2015.
[3] G.B. Huang and L. Chen, "Enhanced random search based incremental extreme learning machine", Elsevier Journal on Neurocomputing, vol. 71, no. 16-18, pp. 3460-3468, 2008.
[4] Zadeh Lotfi, "Knowledge Representation in Fuzzy Logic", IEEE Transactions on Knowledge and Data Engineering, vol. 1, no. 1, pp.89-100, March 1989.
[5] Kindie Biredagn Nahato, Khanna H. Nehemiah and A. Kannan, "Hybrid approach using fuzzy sets and extreme learning machine for classifying clinical datasets", Elsevier Journal on Informatics in Medicine Unlocked, pp.1-11, 2016.
[6] Muhammad Waqar Aslam, Zhechen Zhu and Asoke Kumar Nandi, "Feature generation using genetic programming with comparative partner selection for diabetes classification", Elsevier Journal on Expert system with application, pp. 5402-5412, 2013.
[7] Ren Diao, Fei Chao, Taoxin Peng, Neal Snooke and Qiang Shen, "Feature Selection Inspired Classifier Ensemble Reduction", IEEE Transaction on Cybernetics, vol. 44, no. 8, pp 1259-1268, August 2014.
[8] Xing Wu, Pawe Rzycki and Bogdan M. Wilamowski, "A Hybrid Constructive Algorithm for Single-Layer Feedforward Networks Learning", IEEE Transaction on Neural Networks Learning Systems, vol. 26, no.8, pp 1659-1668, August 2015.
[9] W. H. Teukolsky, S. H. Vetterling and W. T. Flannery,"Numerical recipes in C: The art of scientific computing", Cambridge University Press, 1992.
[10] Huang GB, Zhu QY and Siew CK, "Extreme learning machine: theory and applications", Elsevier Journal on Neurocomputing, pp. 489-501, 2006.
[11] C. V. Subbulakshmi and S. N. Deepa, "Medical Dataset Classification: A Machine Learning Paradigm Integrating Particle Swarm Optimization with Extreme Learning Machine Classifier", Journal Hindawi Publishing Corporation, pp. 1-12, 2015.
[12] Shen Yuong Wong, Keem Siah Yap, Hwa Jen Yap, Shing Chiang Tan and Siow Wee Chang, "On Equivalence of FIS and ELM for Interpretable Rule-Based Knowledge Representation", IEEE Transaction on Neural Network and Learning Systems, vol. 26, no. 7, pp. 1417-1430, July 2015.
[13] K. S. Yap, C. P. Lim and J. Mohamad-Salleh, "An enhanced generalized adaptive resonance theory neural network and its application to medical pattern classification", Journal of Intelligent and Fuzzy System, vol. 21, no. 12, pp. 65-78, 2010.
[14] A. Quteishat and C. P. Lim, "A modified fuzzy minmax neural network with rule extraction and its application to fault detection and classification", Elsevier Journal on Application Soft Computing, vol. 8, no. 2, pp. 985-995, 2008.
[15] K. Polat, S. Gunes, and A. Arslan, "A cascade learning system for classification of diabetes disease: generalized discriminant analysis and least square support vector machine", Elsevier Journal on Expert Systems with Applications, vol. 34, no. 1, pp. 482-487, 2008.
[16] H. Temurtas, N. Yumusak and F. Temurtas, "A comparative study on diabetes disease diagnosis using neural networks", Elsevier Journal on Expert Systems with Applications, vol. 36, no. 4, pp. 8610-8615, 2009.
[17] D.C. alisir and E. Dogantekin, "An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier", Elsevier Journal on Expert Systems with Applications, vol. 38, no. 7, pp. 8311-8315, 2011.
[18] P. Jaganathan and R. Kuppuchamy, "A threshold fuzzy entropy based feature selection for medical database classification", Elsevier Journal on Computers in Biology and Medicine, vol. 43, no. 12, pp. 2222-2229, 2013.
[19] Varma KV, Rao AA, Lakshmi TSM and Rao PN, "A computational intelligence approach for a better diagnosis of diabetic patients", Elsevier Journal on Computer Electrical Engineering , vol. 40, no. 5, pp. 1758-1765, 2014.
[20] Seera M and Lim CP, "A hybrid intelligent system for medical data classification", Elsevier Journal on Expert System Application, vol. 41, no. 5, pp. 2239-2249, 2014.

[21] Mattila J, Koikkalainen J, Virkki A, Van Gils M and Lotjonen J, "Design and application of a generic clinical decision support system for multiscale data", IEEE Transaction on Biomedical Engineering, vol. 59, no. 1, pp. 234-240, January 2012.
[22] F. J. Martnez-Estudillo, C. Hervas-Martnez, P. A. Gutierrez and A. C. Martnez-Estudillo, "Evolutionary product unit neural networks classifiers", Elsevier Journal on Neurocomputing, vol. 72, no. 3, pp. 548-561, 2008.
[23] C. Hervas-Martnez, F. J. Martnez-Estudillo and M. Carbonero-Ruz, "Multilogistic regression by means of evolutionary product-unit neural networks", Elsevier Journal on Neural Networks, vol. 21, no. 7, pp. 951-961, 2008.
[24] Lichman M. 2013 UCI machine learning repository. Irvine, CA: University of Cali-fornia, School of Information and Computer Science, http://archive.ics. uci.edu.
[25] Wenjun Xia, Yoshio Mita and Tadashi Shibata," A Nearest Neighbor Classifier Employing Critical Boundary Vectors for Efficient On-Chip Template Reduction", IEEE Transaction on Neural Networks and Learning Systems, vol. 27, no. 5, pp. 1094-1107, May 2015.
[26] Mahardhika Pratama, Sreenatha G., Anavatti Plamen, P. Angelov and Edwin Lughofer, "PANFIS: A Novel Incremental Learning Machine", IEEE Transaction on Neural Networks and Learning Systems, vol. 25 , no. 1, pp. 55-68, January 2014 .
[27] Rui Zhang, Yuan Lan, Guang-Bin Huang and Zong-Ben Xu, "Universal Approximation of Extreme Learning Machine with Adaptive Growth of Hidden Nodes", IEEE Transaction on Neural Networks and Learning Systems, vol. 23 , no. 2, pp. 365-371, February 2012.
[28] Z. Bai, G.B. Huang, D. Wang, H. Wang and M. B. Westover, "Sparse extreme learning machine for classification", IEEE Transaction on Cybernetics, vol. 44, no. 10, pp. 1858-1870, October 2014.
[29] Whye Loon Tung and Chai Quek, "eFSM_A Novel Online Neural-Fuzzy Semantic Memory Model", IEEE Transaction on Neural Networks , vol. 21, no. 1, pp 136- 157, January 2010.

AUTHORS PROFILE

Archana Kale born in India in 1980 received the B.E degree from Shivaji University, Kolhapur, in 2002 and the M.E. degree in Computer Engineering from Savitribai Phule Pune University, Maharashtra, India, in 2010. She is currently pursuing the Ph.D. degree in Computer Science and Engineering in Shivaji University. She worked as an Assistant Professor, Savitribai Phule Pune University, in Department of Computer Engineering in 2010 and as a Lecturer, Shivaji University, in the Department of Computer Science and Engineering (CSE) in 2002. She is currently a Research Scholar in CSE department of Walchand College of Engineering, Sangli. Her current research interests include Machine Learning, Pattern Recognition and Computer Vision.

Dr. Shefali Sonavane is working as Associate professor and heading Department of Information Technology, Walchand college of engineering, Sangli. She received research funds from DST and AICTE for various projects, supporting work in the area of Computer Vision and Information Security. She has few IPR credentials at her account and extending her research interest further in the fields of Machine Learning and Big Data.

ISBN: 978-81-931456-1-8, Dec. 19-20, 2016

Accelerating Learning Performance of Facial Expression Recognition Using Deep Learning

Umesh Chavan, ubc walchand@rediffmail.com
Walchand College of Engineering ,Sangli,Maharashtra
Dinesh Kulkarni, d b kulkarni@yahoo.com
Walchand College of Engineering ,Sangli,Maharashtra

Abstract: We propose a deep learning based approach using Convolution Neural Networks (CNNs). The goal is to classify each facial image into one of seven (e.g., happy, sad, disgust, surprise, angry, fear expressions) facial emotion categories. Deep Convolution neural network (DCNN) with multiple sequences of layers and appropriate settings is an efficient machine learning method. We trained deep model using Kaggle data set. We developed model in Theano and used Graphics Processing Unit (GPU) computation in order to accelerate the training process. The proposed architecture achieves 61% accuracy with Kaggle data-set. GPU has a speedup of approximately 5 times. GPU has great potential as high-performance co-processor.

ISBN: 978-81-931456-1-8, Dec. 19-20, 2016

Smart Street Management And Emergency System

Nikhil Chaudhary, nikchaudhary101@gmail.com
Pooja Yadav, yadavpooja0217@gmail.com
Aditya Prakash, adityaprakash13@rediffmail.com
Sudiksha Pandey, sudiksha.pandey0@gmail.com
Arti Mohanpurkar, hodcompsoet@dypic.in
Department of Computer Engineering
Dr. D Y Patil School of Engineering & Technology Pune, India

Abstract- This paper describes about a smart street system. This system is an integration of various modules namely, the accident detection, license plate recognition for vehicles emitting smoke, Fire detection in case of accidents and face recognition for the individuals who are missing or wanted as per the database. The first phase of this system deals with the detection and recognition using image processing and pattern recognition while the second phase deals with the reporting of these events for service from the respective facilitation centers.

Keywords: Filtering, Color, Signal processing, Convolution neural network, Metrics, Metafiles, Image processing ,Routing protocols and layout, Gray values, Computer vision, Server, Clustering algorithms, Video Analysis, Calibrated camera, Error Analysis.

Survey on the Prospective and Significance of Deep Learning in Big Data Analysis

Arifa J.Shikalgar, QIP Ph.D. Research Scholar, shikalgar.arifa@walchandsangli. ac.in
Shefali Sonavane, Associate Professor, shefali.sonavane@walchandsangli.ac.in
Information Technology Department,Walchand College of Engineering, Sangli

Abstract— nowadays Big Data describes the strikingly increasing growth in raw data. This data is heterogeneous originated from various sources and having various structures. Most of the data is unstructured and analytics of such unstructured, structured, semi-structured data by using traditional mining algorithm is very crucial. As traditional mining algorithm requires uploading of whole data in a memory for processing and also they are not scalable to process huge complex data. In this paper, a survey on Deep Learning approach for the analysis of Big Data with challenges and future application is discussed. The human brain is capable of processing complex input data; Deep Learning tries to impersonate the same characteristic. Deep Learning model based on high-level abstraction on data. Deep Learning has already proved its ability in speech recognition, computer vision, and natural language processing but it is yet rarely applied on high-dimensional heterogeneous data. Hence, it becomes necessary to check the performance, feasibility of Deep Learning for mining heterogeneous data.

Index Terms— Deep learning, Big Data, heterogeneous data, unstructured data.

REFERENCES

[1]. X. Wu, X. Zhu, Gong-Qing Wu and W. Ding, "Data Mining with Big Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014.
[2]. I. A. Targio Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan, "The rise of "Big Data" on cloud computing: Review and open research issues", Elsevier, Information Systems, Volume 47, Pages 98-115, January 2015.
[3]. Qingchen Zhang, Laurence T. Yang and Zhikui Chen, "Deep Computation Model for Unsupervised Feature Learning on Big Data", IEEE Transactions On Services Computing, Vol. 9, No. 1, 2016.
[4]. Eric P. Xing, Qirong Ho, Wei Dai, et. al., "Petuum: A New Platform for Distributed Machine Learning on Big Data", IEEE Transactions On Big Data, Vol. 1, No. 2, April-June 2015.
[5]. T. Preis, H. S. Moat and E. H. Stanley, "Quantifying trading behavior in ?financial markets using Google trends", Sci. Rep., vol. 3, p. 1684, 2013.
[6]. H. Choi and H. Varian, "Predicting the present with Google trends," Econ. Rec., vol. 88, no. s1, pp. 2–9, 2012.
[7]. S. Lee, J. K. Kim, X. Zheng, Q. Ho, G. Gibson and E. P. Xing, "On model parallelism and scheduling strategies for distributed machine learning", In Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 2834–2842.
[8]. Xue-Wen Chen,"Big Data Deep Learning : Challenges and perspectives" , IEEE Access practical innovations: open solutions,vol 2,pp. 514-525,2014.
[9]. J Gantz, D Reinsel, Extracting value from chaos (EMC, Hopkinton, 2011).
[10]. D Che, M Safran, Z Peng, From Big Data to Big Data mining: challenges, issues and opportunities, in Proceedings of the 18th International Conference on DASFAA (Wuhan, 2013), pp. 1–15.
[11]. H Hu, Y Wen, T Chua, X Li, Toward scalable systems for Big Data analytics: a technology tutorial. IEEE Access 2, 652–687 (2014).
[12]. J Manyika, M Chui, B Brown, J Bughin, R Dobbs, C Roxburgh, AH Byers, Big Data: the next frontier for innovation, competition and productivity (McKinsey Global Institute, USA, 2011).
[13]. Q Wu, G Ding, Y Xu, S Feng, Z Du, J Wang, K Long, Cognitive internet of things: a new paradigm beyond connection. IEEE Internet Things J 1(2), 129–143 (2014).
[14]. CW Tsai, CF Lai, MC Chiang, LT Yang, Data mining for internet of things: a survey. IEEE Commun Surv Tut 16(1), 77–97 (2014).
[15]. A Imran, A Zoha, Challenges in 5G: how to empower SON with Big Data for enabling 5G. IEEE Netw 28(6), 27–33 (2014).
[16]. X Wu, X Zhu, G Wu, W Ding, Data mining with Big Data. IEEE Trans Knowl Data Eng 26(1), 97–107 (2014).
[17]. A Rajaraman, JD Ullman, Mining of massive data sets (Cambridge University Press, Oxford, 2011).
[18]. S Russell, P Norvig, Artificial intelligence: a modern approach (Prentice-Hall, Englewood Cliffs, 1995).
[19]. V Cherkassky, FM Mulier, Learning from data: concepts, theory and methods (John Wiley & Sons, New Jersey, 2007).
[20]. TM Mitchell, The discipline of machine learning (Carnegie Mellon University, School of Computer Science, Machine Learning Department, 2006).
[21]. CM Bishop, Pattern recognition and machine learning (Springer, New York, 2006).
[22]. B Adam, IFC Smith, F Asce, Reinforcement learning for structural control. J Comput Civil Eng 22(2), 133–139 (2008).
[23]. J Langford, Tutorial on practical prediction theory for classification. J Mach Learn Res 6(3), 273–306 (2005).
[24]. Yoshua Bengio, Tutorial on Deep Learning for Big Data (2013)
[25]. Youngmin Cho and Lawrence K Saul. Kernel Methods for Deep Learning. pages 1-9.
[26]. R. Raina, A. Madhavan, and A. Ng, ``Large-scale deep unsupervised learning using graphics processors," in Proc. 26th Int. Conf. Mach.Learn., Montreal, QC, Canada, 2009, pp. 873_880.
[27]. A. Coats, B. Huval, T. Weng, D. Wu, and A. Wu, ``Deep Learning with COTS HPS systems," J. Mach. Learn. Res., vol. 28, no. 3, pp. 1337_1345, 2013.
[28]. Q. Le et al., ``Building high-level features using large scale unsupervised learning," in Proc. Int. Conf. Mach. Learn., 2012.
[29]. J. Dean et al., ``Large-scale distributed deep networks," in Proc. Adv.NIPS, 2012, pp. 1232_1240.
[30]. A. Krizhevsky, I. Sutskever, and G. Hinton, ``ImageNet classification with deep convolutional neural networks," in Proc. Adv. NIPS, 2012, pp. 1106_1114.
[31] Xue-Wen Chen1 , Xiaotong Lin, "Big Data Deep Learning: Challenges and Perspectives," IEEE Access, practical Innovation, open solution, volume 2 ,2014

AUTHORS PROFILE

Arifa J.Shikalgar received the Bachelor's degree in Information Technology from Shivaji University, Kolhapur, Maharashtra. Master's degree in Computer Science and Engineering from Walchand College of Engineering, Sangli, Maharashtra. She is currently perusing Ph.D. in the Computer Science & Engineering. She is having total 11 years' experience. Her field of interest is Artificial Intelligence, and Image processing. She is awarded by Collector Sangli & Ministries for "Innovative Projects" in Feb.2016.

Dr. Shefali Sonavane is working as Associate professor and heading Department of Information Technology, Walchand college of engineering, Sangli. She received research funds from DST and AICTE for various projects, supporting work in the area of Computer Vision and Information Security. She has few IPR credentials at her account and extending her research interest further in the fields of Machine Learning and Big Data.

Addressing Presence of Selfish Nodes in MANET Using Trust

Sandeep A. Thorat, Ph.D. Research Scholar, Shivaji University, Kolhapur, Maharashtra, India
sathorat2003@gmail.com
Dr. P. J. Kulkarni, Professor, CSE Department, Walchand College of Engineering, Sangli, Maharashtra, India
pjk_walchand@rediffmail.com

ABSTRACT: MANET routing protocols are prone to various types of denial of service attacks. These protocols work on basic principle that, all participating nodes are honest and cooperative. However this can't be guaranteed all the time [1]. The nodes participating in a network may exhibit selfish behavior. The packet dropping attack by selfish nodes degrade performance of the network radically [3]. Trust is useful for decision making in many applications. Trust based security mechanisms have edge over cryptographic approaches while dealing with behavioral attacks [2]. The research work studies known literature on using trust in MANET routing and key challenges in it [4]. The ORPSN [5] is trusted opportunistic routing protocol. ORPSN chooses trustworthy nodes closer to the destination for packet forwarding. ORPSN packet delivery ratio is 10% more than CORMAN protocol.

Keywords
MANET, Routing Protocol, Security, Trust, Opportunistic Routing.

REFERENCES

[1] Kannhavong, Bounpadith, Hidehisa Nakayama, Yoshiaki Nemoto, Nei Kato, and Abbas Jamalipour. 2007. "A survey of routing attacks in mobile ad hoc networks." IEEE Wireless Communications 14, no. 5, 85-91.
DOI= http://dx.doi.org/10.1109/MWC.2007.4396947.

[2] Yu, Han, Zhiqi Shen, Chunyan Miao, Cyril Leung, and Dusit Niyato. 2010. "A survey of trust and reputation management systems in wireless communications" Proceedings of the IEEE , no. 10, 1755-1772.
DOI= http://dx.doi.org/10.1109/JPROC.2010.2059690.

[3] Djahel, Soufiene, Farid Nait-Abdesselam, and Zonghua Zhang. 2011. "Mitigating packet dropping problem in mobile ad hoc networks: proposals and challenges." IEEE communications surveys & tutorials 13, no. 4, 658-672.
DOI= http://dx.doi.org/10.1109/SURV.2011.072210.00026.

[4] Thorat, Sandeep A., and P. J. Kulkarni. 2014. "Design issues in trust based routing for MANET." In 5th IEEE International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2014, pp. 1-7.
DOI= http://dx.doi.org/10.1109/ICCCNT.2014.6963101.

[5] Thorat, Sandeep A., and P. J. Kulkarni. 2015. "Opportunistic routing in presence of selfish nodes for MANET." Springer Journal on Wireless Personal Communications 82, no. 2, 689-708.
DOI= http://dx.doi.org/10.1007/s11277-014-2247-4.

AUTHORS PROFILE

Sandeep A. Thorat completed B.E. from Shivaji University Kolhapur. He completed M.Tech. from IIIT Hyderabad with specialization in Information Security. His area of interest is Wireless Networks and Security. He has authored a book on C programming. Presently he is Ph.D. scholar in Walchand College of Engineering, Sangli India.

Overview on Scheduling of Virtual Resources for Scientific Workflows across Federated Cloud

Shivananda R Poojara, shivananda.poojara@ritindia.edu
N V Dharwadkar, nagaraj.dharwadkar@ritindia.edu
Dept. of Computer Sci. and Engineering Rajarambapu Institute of Technology Sakhrale, Sangli Dist.

***Abstract*—** Scientific applications require large computing power, traditionally exceeding the amount that is available within the premises of a single institution. Scientific experiments are being conducted in collaboration with teams that are dispersed globally. Each team shares its data and utilizes distributed resources for conducting experiments. As a result, scientific data are replicated and cached at distributed locations around the world. These data are part of application workflows, which are designed for reducing the complexity of executing and managing on distributed computing environments. Cloud computing is emerging as a viable platform for scientific exploration. Workflow execution cannot be achieved only by using public clouds or private, because of cost and time factor. In scientific workflow execution resources are to be leased dynamically to fulfill the request in time and cost efficient. So there is a combination of multiple clouds required to fulfill the request. Objective of research is to study and analysis of dynamic provisioning broker policies for multi cloud environments.

Keywords: Scientific Workflow, Cloud Computing, federated cloud, Dynamic provisioning.

REFERENCES

[1] Fuhui Wu QingboWu,Yusong Tan "Workflow scheduling in cloud: a survey", © Springer Science+Business Media New York 2015
[2] Chiaki Sato,Luke M. Leslie,YoungChoon Lee "Running Data-Intensive Scientific Workflows in the Cloud", International Conference on Parallel and Distributed Computing, Applications and Technologies
[3] Gideon Juve and EwaDeelman, "ScientificWorkflows in the Cloud", University of Southern California, Marina del Rey, CA e-mail: juve@usc.edu
[4] L. Wu, S. Panday, R. Buyya, "A Particle Swarm Optimization-Based Heuristic for Scheduling Workflow Applications in Cloud Computing Environments", proceedings of the IEEE International Conference on Advanced Information Networking and Applications.
[5] "A Survey on Deadline Constrained workflow scheduling in cloud environment", Nallakumar. R1, SruthiPriya. K. S2
[6] X. Liu et al., The Design of Cloud Workflow Systems,SpringerBriefs in Computer Science,"Cloud Workflow System Architecture"
[7] MarekWieczorek, RaduProdan and Thomas Fahringer, "Scheduling of Scientific
[8] Gideon Juve and EwaDeelman, "ScientificWorkflows in the Cloud"
[9] Arabnejad H, Barbosa JG (2014) A budget constrained scheduling algorithm for workflow applications.J Grid Comput, pp 1–15
[10] R. Prodan and T. Fahringer, "Dynamic Scheduling of Scientific Workflow Applications on the Grid using aModularOptimisation Tool: A Case Study", In 20th Symposium of Applied Computing (SAC 2005), Santa Fe,New Mexico, USA, March 2005. ACM Press.
[11] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob, and D. S. Katz. Pegasus: A framework for mapping complex scientific workflows onto distributed systems. Sci. Program., 13(3):219–237, 2005.
[12] A. Salman. Particle swarm optimization for task assignmentproblem. Microprocessors and Microsystems, 26(8):363–371, November 2002.
[13] W.N. Chen, J. Zhang and Yang. Y,"Workflow Scheduling in Grids: An Ant Colony Optimization Approach." Evolutionary Computation. IEEE Conference, Pg. 3308-3315, September 2007.
[14] Jai YU and RajkumarBuyya, "Scheduling Scientific Workflow Applications with Deadline and Budget Constraints using Genetic Algorithms". Journal Scientific ProgrammingScientific workflows, Vol 14, December 2006.
[15] Gencyilmaz. A particle swarm optimization algorithm for makespan and total flowtime minimization in the permutation flowshop sequencing problem. European Journal of Operational Research, 177(3):1930–1947, March 2007.
[16] Juve G, Deelman E, Vahi K, Mehta G, Berriman B, Berman BP, Maechling P (2010) Data sharing options for scientific workflows on amazon ec2. In: Proceedings of the 2010 ACM/IEEE international conference for high performance computing, networking, storage and analysis, IEEE Computer Society, pp 1–9

Blink detection using Electrooculography for stress detection

Mrs.Mamata S.Kalas
M.Tech (Computer Sci. and Engg.), Phd(Persuing), KIT's College of Engg,Kolhapur
Dr. B. F. Momin, PhD. (Computer Sci & Engg.),
Associate Professor,Department of CSE, Walchand College of Engg,Sangli

Abstract- There are several human activities where the awareness and conscious control is a very important factor. The research work has significance in the general field of safety of individual working in monotonous environment like, academic work, data operators with heavy equipment operation, hazardous materials manipulation. Excessive mental workload has negative impact on performance and can lead to errors with disastrous outcome. Therefore developing systems for monitoring the mental workload of operator and reducing the stress of the operator when he is not paying adequate attention to the performance of task is essential in order to prevent system failures.

We have to characterize mental states of operator performance, by finding patterns in timely changing physiological, measures like EOG(ElectroOcculography), with eye blinks and EEG signals.

In this work, suppression of blink artifacts is considered. There are two approaches in blink effect suppression. The first approach is the independent processing of EEG channels like subtracting the filtered electrooculogram (EOG) channel from EEG or using wavelet transform to remove blink artifacts. The second approach uses the correlation of blink functions from different channels. Blink waveform in electrooculogram (EOG) data is used to develop and adjust the method of stress detection in operators and academicians who are engaged in continuous reading and learning operations. We devise different features based on these characteristics and select a subset of them using feature selection. We validate the method using dataset RMS Maximus. We have implemented Blink detection algorithm which is used to study the EOG signal and give explanation of different kind of waveforms in EOG signal, and give suggestions to improve the blink detection algorithm.

Keywords: Eye tracking, Electoocculogram(EOG), Blink detection algorithm , Electroencephalograph (EEG).

REFERENCES

[1] Andreas Bulling, Student Member, IEEE, Jamie A. Ward, Hans Gellersen, and Gerhard Tro¨ ster, Senior Member, IEEE, Eye Movement Analysis for Activity Recognition Using Electrooculography, IEEE Transactions on pattern analysis and machine intelligence,Vol 33,No 4, April 2011
[2] Mu Li, Jia-Wei Fu and Bao-Liang Lu, "Estimating Vigilance in Driving Simulation using Probabilistic PCA", 30th Annual International IEEE EMBS Conference Vancouver, British Columbia, Canada, August 20-24, 2008 Electric Engineering and Computer August 19-22, 2011, Jilin, China
[3] Weirwille,W.W. (1994), "Overview of Research on Driver Drowsiness Definition and Driver Drowsiness Detection," 14th International Technical Conference on Enhanced Safety ofVehicles, pp 23-26.
[4] QiangJi, PeilinLan, Carl Looney "A Probabilistic Framework for Modeling and Real-Time Monitoring Human Fatiguel, IEEE Transactions on Systems, Man, and Cybernetics—part A: Systems and Humans, vol. 36, no. 5, September 2006
[5] Li Shiwu, Wang Linhong, Yang Zhifa, JiBingkui, QiaoFeiyan, Yang Zhongkai, "An Active Driver Fatigue Identification Technique Using Multiple Physiological Features" ,2011 International Conference on echatronic Science
[6] Li-Chen Shi, Bao-Liang Lu, "Dynamic Clustering for Vigilance Analysis Based on EEG", Proceedings of International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 54-57, Vancouver, BC Canada, Aug. 2008.
[7] Dhaval Pimplaskar et al Real Time Eye Blinking Detection and Tracking Using Opencv Int. Journal of Engineering Research and Applications ,ISSN : 2248-9622, Vol. 3, Issue 5, Sep-Oct 2013, pp.1780-1787.
[8] Anund A, Kecklund G, Peters B, Forsman A, Lowden A, Akerstedt T. (2008). Driverimpairment at night and its relation to physiological sleepiness. Scandinavian Journal of Work, Environment and Health. 34(2):142-150.
[9] Zafer Savas, Real time detection and tracking of human eyes in video sequences, A thesis submitted to School of Natural and applied sciences of middle East Technical university. Aug. 2005
[10] Borah, J. (2006). Eye movement measurement techniques for Encyclopedia of Medical Devicesand Instrumentation. John Wiley & Sons, Inc. p263
[10] Jammes B, Sharabty H.,& Esteve D. (2008). Automatic EOG analysis: A first step toward automatic drowsiness scoring during wake-sleep transitions. Somnologie – Schlafforschung und Schlafmedizin,Vol. [11],.Number 3, 227-232, DOI: 10.1007/s11818-008-0351-y
[12] Evinger,C., Manning, K., & Sibony ,P. (1991). Eyelid Movements Mechanisms and Normal Data. Investigative Ophthalmology & Visual Science, Vol. 32: 387-400.
[13] Fogarty, C., & Stern, J. (1989). Eye movements and blinks: their relationship to higher cognitive processes. International Journal of Psychophysiology Vol. 8, Issue 1, September, 35 - 42.
[14] Johns M W, Tucker A (2005). The amplitude-velocity ratios of eyelid movements during blinks:changes with drowiness. Journal of Sleep Research, 28:A122
[15] Johns M W, Tucker A, Chapman R, Crowley K, Micheal N (2007). Monitoring eye and eyelid movements by infrared reflectance oculography to measure drosiness in drivers. Journal of Sleep Research, 11:234-242. Doi 10.1007/s11818. 007. 0311.
[16] Iwasaki, M., & Kellinghaus, C. (2005). Effects of eyelid closure, blinks, and eye movements on the electroencephalogram. Clinical Neurophysiology 116, 878 - 885.
[17] Kircher, A. (2001). General Information Vitaport II. Linköping: VTI (Swedish National Road and Transport Research Institute).
[18] Malmivuo, J., & Plonsey, R. (1995). Principles and Applications of Bioelectric and Biomagnetic Fields. New York, Oxford, Oxford University Press, Inc. National Sleep Foundation. Facts and Stats http://drowsydriving.org/about/facts-and-stats/NTSB, (1999).
[19] Evaluation of U.S. Department of Transportation: efforts in the 1990s to address operation fatigue, Report No. Safety Report NTSB/SR-99/01, National Transportation Safety Board, Washington, D.C.

[20] Robert D. Peters., Esther Kloeppel., Elizabeth Alicandri,. & Jean E. Fox, (1999). Effects of Partial and Total Sleep Deprivation On Driving Performance, US Department of Transportation, February 1999.
[21] Rottach K G, Das V E, Wohlgemuth W, Zivotofsky A Z., & Leigh R J. (1998). Properties of Horizontal Saccades Accompanied by Blinks, Neurophysiol June 1, 79:2985-2902.
[22] Thorslund, B. (2003). Electrooculogram Analysis and Development of a System for Defining Stages of Drowsiness. Linköping University, Linköping.
[23] Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A., & Kliegl, R. (2011). Coregistration of eye movements and EEG in natural reading: Analyses & Review. Journal of Experimenta Psychology: General, 140 (4), 552-572 [plugin reference paper, PDF]
[24] Engbert, R., & Mergenthaler, K. (2006). Microsaccades are triggered by low retinal image slip. PNAS, 103 (18), 7192-7197
[25] Plöchl, M., Ossandon, J.P., & König, P. (2012). Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. Frontiers in Human Neuroscience, doi: 10.3389/fnhum.2012.00278
[26] Dimigen, O., Valsecchi, M., Sommer, W., & Kliegl, R. (2009). Human microsaccade-related visual brain responses. J Neurosci, 29, 12321-31
[27] Dimigen, O., Kliegl, R., & Sommer, W. (2012). Trans-saccadic parafoveal preview benefits in fluent reading: a study with fixation-related brain potentials. Neuroimage, 62 (1), 381-393

AUTHORS PROFILE

Mrs. Mamata S. Kalas , Having been graduate from University Of Mysore, in 1993, From B.I.E.T,Davangere, started her professional carrier there itself. Since 1995, she has worked as lecturer at D.Y.Patil's college of Engg., Bharati Vidyapeeth's College Of Engg.,Kolhapur.She is M.Tech(CST) Graduate and her dissertation work is based on image segmentation using parametric distributional clustering. She has been awarded with M.TECH (CST) from Shivaji University, Kolhapur of Maharashtra in June 2009. She is persuing Ph. D in computer science and Engineering at walchand College of Engineering, Reseach center, Shivaji University, under the guidance of Dr.B.F.Momin. She is currently working as an Associate Professor at KIT'S College of Engg, Kolhapur.She is in her credit, 19 Years of teaching experience, eight papers presented for international conferences,6 papers presented for n national conferences, seven papers published in international journals. Her areas of interest are pattern recognition and artificial intelligence, Computer Architecture, System programming.

Dr. Bashirahamad F. Momin is working as Associate Professor & Head, Dept. of Computer Science & Engineering, Walchand College of Engineering Sangli, Maharashtra State India. He received the B.E. and M.E. degree in Computer Science & Engineering from Shivaji University, Kolhapur, India in 1990 and 2001 respectively. In February 2008, he had completed his Ph.D. in Computer Science and Engineering from Jadavpur University, Kolkata. He is recognized Ph.D. guide in Computer Science & Engineering at Shivaji University, Kolhapur. His research interest includes pattern recognition & its applications, data mining and soft computing techniques, systems implementation on the state of art technology. He was a Principal Investigator of R & D Project titled "Data Mining for Very Large Databases" funded under RPS, AICTE, New Delhi, India. He had delivered a invited talk in Korea University, Korea and Younsea University, Korea. He had worked as "Sabbatical Professor" at Infosys Technologies Ltd., Pune. He is a Life Member of "Advanced Computing and Communication Society", Bangalore INDIA. He was a student member of IEEE, IEEE Computer Society. He was a member of International Unit of Pattern Recognition and Artificial Intelligence (IUPRAI) USA.

Parallel Implementation of Bellman-Ford Algorithm Using OpenMP

Ganesh G. Surve, surveganesh93@gmail.com
Medha A. Shah, medha.shah@walchandsangli.ac.in
Department of Computer Sci. and Engineering, Walchand Collage of Engineering, Sangli (M.S), India

Abstract— To find out the shortest paths from a single source to all other vertices is a common problem in graph theory. The Bellman-Ford's algorithm is the solution that solves such a single-source shortest path (SSSP) problem in large and complex graph set and it is better suited to be parallelized for many thread architectures. The drawback of sequential Bellman-Ford algorithm is that it is a very time-consuming process for the large number of vertices in a graph. Parallel implementation of Bellman-Ford's algorithm significantly reduces the number of relaxing operation as well as a number of iterations and time required for search the shortest path is minimized. This paper represents a parallel implementation of the Bellman-Ford algorithm based on Pthread using OpenMP (Open Multi-Processing). OpenMP based parallel computing is an alternate way to increase the speed up for search task. This paper gives detailed information about modified parallel implementation of Bellman-Ford's algorithm to solve optimization problem and to improve the work efficiency and performance of GPU-based computing.

Keywords: Bellman-Ford, SSSP, GPU (Graphics Processing Unit), OpenMP (Open Multi-Processing).

REFERENCES

[1] F. Busato and N. Bombieri, "An efficient implementation of the bellman-ford algorithm for Kepler GPU architectures," IEEE Trans. On Parallel and Distributed Systems, vol. 27th, no. 8, pp. 1{13, 2016}.
[2] P. Harish and P. Narayanan, "Accelerating large graph algorithms on the GPU using CUDA," in Proc.Conference. High Perform. Comput, vol. 14, no. 4, pp. 197 {208, Dec. 2007.
[3] A. Davidson, S. Baxter, M. Garland, and J. Owens, "Work-efficient parallel GPU methods for single-source shortest paths," 2014, pp.349–359.
[4] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. Problem 24-1: Yen's improvement to Bellman-Ford. Introduction to Algorithms, 2nd edition, pp. 614-615.MIT Press, 2001.
[5] U. Meyer and P. Sanders, "Δ-stepping: a parallelizable shortest path algorithm," Journal of Algorithms, vol. 49, no. 1, pp. 114–152, 2003.
[6] P. J. Martin, R. Torres, and A. Gavilanes, "CUDA solutions for the SSSP problem," in Proceedings of the 9th International Conference on Computational Science: Part I, ser. ICCS '09, 2009, pp. 904–913.
[7] AydınBuluc, John R. Gilbert and CerenBudak, "Solving Path Problems on the GPU" Journal Parallel Computing Volume 36 Issue 5-6, June 2010 Pages 241-253.
[8] J. Y. Yen., An algorithm for finding shortest routes from all Source nodes to a given destination in general networks Quarterly of Applied Mathematics 27:526-530, 1970.
[9] Pankhari Arawal and Maitreyee Dutta, "New Approach of the bellman-ford algorithm on GPU using CUDA," International Journal of Computer Application , vol. 110 th, no. 13, Jan 2015.
[10] Yefim Dinitz, Rotem Itzhak, Hybrid Bellman-Ford-Dijkstra's Algorithm. https://www-m9.ma.tum.de/graph-algorithms/sppbellman-Ford /indexen.htm
[11] http://techieme.in/shortest-path-using-bellman-ford algorithm/
[12] http://faculty.ycp.edu/~dbabcock/PastCourses/cs360/lectures/ lecture21.html

AUTHORS PROFILE

Ganesh G. Surve Received B. Tech (IT) Degree from Shivaji University, Kolhapur in the year 2015. He is currently M.Tech student, at Walchand College of Engineering, Sangli. His area of research is High Performance Computing.

Mrs. Medha A. Shah received M.E (CSE) Degree from Shivaji, University, Kolhapur in year 2008. She is currently Ph.D. student, at Walchand College of Engineering, Sangli. Her area of research is High Performance Computing. At preset she is working as Asst. Professor in CSE Dept. of Walchand College of Engineering, Sangli. Her main research area includes design of parallel algorithms for content based distributed systems. She is author of some papers in the area of parallel and distributed systems, published in international journal and conference proceedings. She has presented her research work in international conferences held at Irvine, USA and Oslo, Norway.

Real time monitoring system for road traffic detection using twitter and spark

Ketan R. Pandhare, ketanpandhare03@gmail.com
Medha A. Shah, medha.shah@walchandsangli.ac.in
Department of Computer Science and Engineering, Walchand College of Engineering, Sangli, India

Abstract: In this age of Information and Communication Technology, people rely on social media to get regular updates about world events and happenings. Sites like Google, Facebook and Twitter contribute a lot to give real time information feeds on smart devices. In every second, thousands of events occur around. It necessitates people to receive relevant information on these sites. This paper presents a real time monitoring system which uses Twitter as a source of information. Here we fetch tweets related to road traffic congestion, car accidents etc. and process those tweets by text mining techniques using apache spark. The goal of this system is to assign a class label to each tweet so as to identify whether it is a traffic related tweet. We use Geo-tags (longitude and latitude) to fetch location wise tweets. For classification model we use support vector machine classifier.

Keywords: Traffic event detection, Apache spark, Twitter4j, tweets classification, text mining.

REFERENCES

[1] G. Anastasi et al."Urban and social sensing for sustainable mobility in smart cities," in Proc. IFIP/IEEE Int. Conf. Sustainable Internet, ICT Sustainability, Palermo, Italy, 2013, pp 1-4.
[2] A. Rosi et al., "Social sensors and pervasive services: Approaches and perspectives," in Proc. IEEE Int. Conf. PERCOM Workshops, Seattle, WA, USA, 2011, pp. 525-530.
[3] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information. Berlin, Germany: Springer-Verlag, 2004
[4] A. Schulz, P. Ristoski, and H. Paulheim, "I see a car crash: Real-time detection of small scale incidents in microblogs,: in The Semantic Web: ESWC 2013 Satellite Events, vol. 7955. Berlin, Germany: Springer-Verlag, 2013, pp. 22-33.
[5] C. Chew and G. Eysenbach, "Pandemics in the age of Twitter: Content analysis of tweets during the 2009 H1N1outbreak," PLoS ONE, vol. 5, no. 11, pp. 1-33 Nov. 2010.
[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweets analysis for real-time event detection and earthquake reporting system development," IEEE Trans. Knowl. Data Eng., vol.25, no. 4, pp. 919-931, Apr. 2013.
[7] M. Krstajic, C. Rohrdantz, M. Hund, and A. Weiler, "Getting there first: Real-time detection of real-world incidents on Twitter" in Proc. 2nd IEEE Work Interactive Vis. Text Anal.—Task-Driven Anal. Soc. Media IEEE VisWeek," Seattle, WA, USA, 2012.
[8] B. De Longueville, R. S. Smith, and G. Luraschi, "OMG, from here, I can see the flames!: A use case of mining location based social networks to acquire spatio-temporal data on forest fires," in Proc. Int. Work. LBSN, 2009 Seattle, WA, USA, pp. 73–80.
[9] R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, "TEDAS: A Twitter based event detection and analysis system," in Proc. 28th IEEE ICDE, Washington, DC, USA, 2012, pp. 1273–1276.
[10] P. Agarwal, R. Vaithiyanathan, S. Sharma, and G. Shro, "Catching the long-tail: Extracting local news events from Twitter," in Proc. 6th AAAI ICWSM, Dublin, Ireland, Jun. 2012, pp. 379–382.
[11] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Twitcident: fighting fire with information from social web streams," in Proc. ACM 21st Int. Conf. Comp. WWW, Lyon, France, 2012, pp. 305–308.
[12] N. Wanichayapong, W. Pruthipunyaskul, W. Pattara-Atikom, and P. Chaovalit, "Social-based traffic information extraction and classification," in Proc. 11th Int. Conf. ITST, St. Petersburg, Russia, 2011, pp. 107–112.
[13] M. Habibi, Real World Regular Expressions with Java 1.4. Berlin, Germany: Springer-Verlag, 2004.
[14] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining," LDV Forum-GLDV J. Comput. Linguistics Lang. Technol., vol. 20, no. 1, pp. 19–62, May 2005.
[15] Y. Zhou and Z.-W. Cao, "Research on the construction and filter method of stop-word list in text preprocessing," in Proc. 4th ICICTA, Shenzhen, China, 2011, vol. 1, pp. 217–221.
[16] W. Francis and H. Kucera, "Frequency analysis of English usage: Lexicon and grammar," J. English Linguistics, vol. 18, no. 1, pp. 64–70, Apr. 1982.
[17] M. F. Porter, "An algorithm for suffix stripping," Program: Electron. Library Inf. Syst., vol. 14, no. 3, pp 130–137, 1980.
[18] L. M. Aiello et al., "Sensing trending topics in Twitter," IEEE Trans. Multimedia, vol. 15, no. 6, pp. 1268–1282, Oct. 2013.
[19] C. Shang, M. Li, S. Feng, Q. Jiang, and J. Fan, "Feature selection via maximizing global information gain for text classification," Knowl.-Based Syst., vol. 54, pp. 298–309, Dec. 2013.
[20] L. H. Patil and M. Atique, "A novel feature selection based on information gain using WordNet," in Proc. SAI Conf., London, U.K., 2013, pp. 625–629.
[21] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995.
[22] G. H. John and P. Langley, "Estimating continuous distributions in Bayesian classifiers," in Proc. 11th Conf. Uncertainty Artif. Intell., San Mateo, CA, 1995, pp. 338–345.
[23] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA, USA: Morgan Kaufmann, 1993.
[24] T. T. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Trans. Inf. Theory, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[25] Eleonora D' Andrea, Pietro Ducange, Beatrice Lazzerini,"Real –Time Detection of Traffic From Twitter Stream Analysis," IEEE Trans. On Intelligent Transportation System, vol 16, No.4 Aug 2015
[26] Apache Spark:http://spark.apache.org/
[27] Twitter4j: http://twitter4j.org/en/index.html.
[28] Twitter: https://twitter.com.
[29] Stop word list:http://snowball.tartarus.org/algorithms/english/stop.txt

[30] J. Platt, "Fast trainning of support vector machines using sequential minimal optimization," in Advances in Kernel Methods: Support Vector Learning, B. Schoelkopf, C. J. C. Burges and A. J. Smola, Eds. Cambridge, MA, USA, MIT Press, 1999,pp.185-208.

AUTHORS PROFILE

Ketan R. Pandhare Received B.E. (CSE) Degree from Shivaji University, Kolhapur in year 2015. He is currently M.Tech student, at Walchand College of Engineering, Sangli. His main area of research is Big data and analytics.

Mrs. Medha A. Shah received M.E (CSE) Degree from Shivaji, University, Kolhapur in year 2008. She is currently Ph.D. student, at Walchand College of Engineering, Sangli. Her area of research is High Performance Computing. At preset she is working as Asst. Professor in CSE Dept. of Walchand College of Engineering, Sangli. Her main research area includes design of parallel algorithms for content based distributed systems. She is author of some papers in the area of parallel and distributed systems, published in international journal and conference proceedings. She has presented her research work in international conferences held at Irvine, USA and Oslo, Norway.

Secured Deep Learning in Cloud- An Approach

Mr. S. S. Sayyad, Dr. D B Kulkarni

Abstract: Deep Learning is an emerging technique which is used to model, classify, cluster and predict complex datasets involving speech, text, images, videos, etc. The accuracy factor involved in Deep Learning Algorithms has made them into the core of emerging Artificial Intelligence based services. Many organizations which involve dealing with huge amount of data on a larger scale have been the core user of this technique since the success has handled the larger volume of training data quite effectively. When data is involved in huge volume for learning algorithms then it surely produces privacy issues. Many data sources preferably from Medical Hospitals and Institutions that are looking for Artificial Intelligence based learning approaches are willing to produce learning models without comprising on the data privacy issues.

In this paper, we propose a practical system that will enable multiple parties to collaboratively learn a neural network model for a common objective without compromising the privacy issues of the data involved. Our proposed system will provide access control based cloud system for users to upload their data in Cloud. Most complex operations of learning algorithms i.e. Activation Function and Update of the model can be outsourced to cloud based on GPU. We propose to execute our secured deep learning in cloud approach on benchmark datasets.

ABOUT AUTHORS

Mr S S Sayyad is a research scholar at Walchand College of Engineering, Sangli. He has completed his M.Tech in Computer Science and Engineering from Walchand College of Engineering, Sangli in 2012 and B.E in Computer Science and Engineering from Karmaveer Bhaurao Patil College of Engineering, Satara in 2007. His areas of interest are High Performance Computing, Cloud Computing. He is currently working as Assistant Professor at Annasaheb Dange College of Engineering and Technology Ashta for last 7 years. He also has 2 years of industrial experience as Software Engineer at Zensar Technologies, Pune.

Dr. D. B. Kulkarni received the doctorate degree in Computer Science and Engineering from Shivaji University, Kolhapur in the year 2005. Currently he is a professor in Information Technology Department of Walchand college of Engineering, where he teaches many courses in the area of Computer Science. During his professional life, he has been involved in several R&D projects funded by AICTE, DRDO and UGC. Recently he received Early Adopter award of National Science Foundation's (NSF) of Technical Committee on Parallel Processing (TCPP) of US$1500 for framing and adapting curriculum on “Parallel Computing” in UG and PG curriculum in the institute. He was Visiting Professor in institute of Computer Technology (ICT) of Vienna University of Technology, Austria, in 2010. He is coauthor of tens of scientific papers published in international journals and conference proceedings. His research interests include the area of High Performance computing and Computer Network.

National Research Symposium on Computing - RSC 2016

The WCE Research Symposium on Computing 2016 has welcomed submissions from doctoral students, research scholars who are working in the area of Computer Science and Engineering. The WCE Research Symposium on Computing 2016 will provide PhD students and research scholars working in all areas of computer science with an unique opportunity to present their work and interact with renowned and experienced researchers in their field. The symposium will support students by providing independent and constructive feedback about their already completed and more importantly, planned research work. The research symposium also aims at establishing a supportive community of doctoral students, research scholars and potential collaborators working in the relevant areas of Computer Science and Engineering. Promoting social and intellectual interactions among students and researchers from academia and industry is one of the motivations behind organization of Research Symposium. The symposium has been participated by prominent experts who would actively participate in critical discussions.

Venue:
Ajit Gulabchand Central Library
Walchand College of Engineering, Sangli.

Email
chair.rsc@walchandsangli.ac.in

Website:
http://rsc.wce.ac.in

ISBN 978-81-931546-1-8
9 788193 154618