Jan Zizka
Brajesh Kumar (Eds)

Computer Science & Information Technology

Second International Conference on Advances in Computer Science and Information Technology (ACSTY 2016)
Chennai, India, November 26~27, 2016

AIRCC Publishing Corporation

Volume Editors

Jan Zizka,
Mendel University in Brno, Czech Republic
E-mail: zizka.jan@gmail.com

Brajesh Kumar K,
IIT-Roorkee, India
E-mail: bkkaushik23@gmail.com

ISSN: 2231 - 5403
ISBN: 978-1-921987-59-5
DOI : 10.5121/csit.2016.61401 - 10.5121/csit.2016.61404

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

The Second International Conference on Advances in Computer Science and Information Technology (ACSTY 2016) was held in Chennai, India, during November 26~27, 2016. The Second International Conference on Natural Language Processing (NATP 2016) was collocated with the ACSTY-2016. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ACSTY-2016, NATP-2016 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, ACSTY-2016, NATP-2016 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the ACSTY-2016, NATP-2016.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

Jan Zizka
Brajesh Kumar

Organization

General Chair

Natarajan Meghanathan	Jackson State University, USA
Dhinaharan Nagamalai	Wireilla Net Solutions, Australia

Program Committee Members

Aijaz	University of Jyväskylä, Finland
Andhe Dharani	R.V.College of Engineering, India
Ankit Chaudhary	Truman State University, USA
Chun-Yi Tsai	National Taitung University,Taiwan
Dabin Ding	University of Central Missouri, Missouri
Debajit Sensarma	University of Calcutta, India
Diptoneel Kayal	West Bengal University of Technology, India
Farhan	University of Indonesia, Indonesia
Foudil Cherif	Biskra University, Algeria
George Dharma Prakash Raj E	Bharathidasan University, India
Grienggrai Rajchakit	Maejo University, Thailand.
Gullanar M Hadi	Salahaddin University, Hawler, Iraq
Hamadouche M	Universite Saad Dahlab de Blida, Algeria
Hamdi M	National Engineering School of Tunis, Tunisia
Iancu Mariana	Bioterra University of Bucharest, Romania
Jacques Epounde Ngalle	Robert Morris University, USA
Jayaraj	Bharathidasan University, India
Koushik Majumder	West Bengal University of Technology, India
Kuppusamy K	Alagappa University, India
Lawrence J. Osborne	Lamar University, USA
Majlinda Fetaji	South East European University, Macedonia
Marcin Michalak	Silesian University of Technology, Poland
Meyyappan T	Alagappa University, India
Mohamed Khamiss	Suez Canal University, Egypt
Murat Topaloğlu	Trakya University, Turkey
Othman Chahbouni	University of Hassan II Casablanca, Morocco
Paramartha Dutta	Visvabharati University, West Bengal
Peiman Mohammadi	Islamic Azad University, Iran
Prachi Ahlawat	ITM University, India
Raed I Hamed	University of Anbar Ramadi, Iraq
Roheet Bhatnagar	Manipal University, India
Saad M.Darwish	Alexandria University, Egypt
Saadat Pourmozafari	Tehran Poly Technique, Iran
Sachin Chirgaiya	Oriental University, India
Sandhya M	B.S.Abdur Rahman University, India
Seyyed AmirReza Abedini	Islamic Azad University, Iran
Shahid Siddiqui	Integral University, India

Shuxiang Xu	University of Tasmania, Australia
Smain Femmmam	UHA University, France
Sneha Thombre	University of Pune, India
Sokyna Qatawneh	Al-Zaytoonah University of Jordan, Jordan
Suryakanthi T	Botho University, Botswana
Taruna S	Banasthali University, India
Timothy Roden	Lamar University, USA
Wei cai	University of Hawaii at Manoa, Hawaii
Willie K Ofosu	Penn State Wilkes-Barre, USA
Yen-Chun Jim Wu	National Taiwan Normal University, Taiwan
Zhiyong Shan	University of Central Missouri, United States

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Digital Signal & Image Processing Community (DSIPC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

The Second International Conference on Advances in Computer Science and Information Technology (ACSTY 2016)

The Second International Conference on Natural Language Processing (NATP 2016)

COMPUTATIONAL METHODS FOR FUNCTIONAL ANALYSIS OF GENE EXPRESSION

Houda Fyad[1], Fatiha Barigou[1], Karim Bouamrane[1]

[1]LIO Laboratory, Department of Computer Science, Faculty of Exact and Applied Sciences University of Oran 1 Ahmed Ben Bella BP 1524, 31000 El M'naouer Oran, Algeria

houdafyad82@gmail.com, fatbarigou@gmail.com, kbouamrane@gmail.com

ABSTRACT

Sequencing projects arising from high throughput technologies including those of sequencing DNA microarrays allowed to simultaneously measure the expression levels of millions of genes of a biological sample as well as annotate and identify the role (function) of those genes. Consequently, to better manage and organize this significant amount of information, bioinformatics approaches have been developed. These approaches provide a representation and a more 'relevant' integration of data in order to test and validate the hypothesis of researchers throughout the experimental cycle. In this context, this article describes and discusses some of techniques used for the functional analysis of gene expression data.

KEYWORDS

Microarray, genes, genome annotation, functional analysis, expression data, datamining, clustering, classification, Gene ontology.

1. INTRODUCTION

The successful developments of high throughput sequencing technology including those of sequencing DNA microarrays generated a large volume of genomic data. The massive data produced presents a significant challenge for data storage and analysis. In this case, bioinformatics tools are essential for data management.

This technology allows to measure the simultaneous expression of a large number of genes, or even all the genes contained in the genome under many and varied conditions. Also, it identifies the rate of gene expression (over or under expressed); characterization of genes differentially expressed; the establishment of a characteristic profile of a given biological state. Therefore, it provides to researchers the opportunity to study the coordinated behavior of genes and so better understanding the function of a gene in an experimental situation.

Jan Zizka et al. (Eds) : ACSTY, NATP - 2016
pp. 01– 14, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61401

Thus, the transition of the genome sequencing to the annotated genome gave rise to methods, tools, and bioinformatics platforms, to help many areas of biology to manage and organize this mass of data. Some of these approaches using data mining have been developed to determine the similar expression profiles of genomic data. Others have used controlled vocabularies or ontologies to capture the semantics of biological concepts describing biological objects such as genomic sequences, genes or gene products. And some have combined the two above-mentioned approaches. All of this, providing biologists with a more "relevant" representation and data integration allowing them to analyze their genomic data, test and validate their assumption throughout the experimental cycle.

This article gives a comprehensive overview of the different approaches employed in the functional analysis of gene expression data. The rest of the paper is organized as follows: section two introduces the concept of the genomic annotation with its three levels of complexity. Section Three describes the different data mining techniques used in functional analysis of gene expression. The fourth section deals with the use of gene ontology to build a gene expression profile. And, finally, the fifth section provides some concluding remarks and gives an outlook for future works.

2. GENOME ANNOTATION

Definition and strategies for genome annotation

Genome annotation (or DNA annotation) is extraction, definition, and interpretation of features on the genome sequence derived by integrating computational tools and biological knowledge. An annotation is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it. Annotation could be:

- gene products names
- functional characteristics of gene products
- physical characteristics of gene/protein/genome
- overall metabolic profile of the organism

For example, genome annotation is notably used by biologists for identification of different genes expressed in plants organs (root, leaf,...) during a cycle of development like the *Arabidopsis thaliana* plant [1.2], also it is used for identification of genes involved in the rice tolerance to salinity [1, 2] and possibly for the discovery of new functions by the association of genes with "known" genes based on the co-expressed and co-regulation in coral [3]. For the Drosophila, it was to determine the present/absent genes in neural flow and synaptic transmission routing [4]. For the mouse, the study consists of analyzing the over or under expression of genes across different genetic manipulation of embryos and adults and the effects of environmental conditions [5]. In medicine [6], it allows distinguishing and classifying types of tumors, knowing the genes expressed on a large number of patients to observe the effect of a drug (e.g. anti-cancer), examine the effect of a treatment on the expression of genes, to compare healthy tissue from diseased tissue, treated against untreated.

The process of annotation can be divided into three levels [7]:

- The syntactic or structural annotation it identifies sequences presenting a biological relevance (genes, signals, repetitions, etc).
- The functional annotation it predicts the potential functions of the previously identified genes (similarities of sequences, patterns, structures, etc) and collects any experimental information (literature, big data sets, etc).
- Relational or contextual annotation it determines the interactions between the biological objects (families of genes, regulatory networks, metabolic networks, etc).

Also, these different levels of annotation are not separated, but intermingle, and are very closely related. The genomic annotation is precisely to interconnect these three different levels [7].
In the next sections, we present methods and techniques using (i) data mining for identification of genes co-expressed in an analysis of expression data. (ii) Ontology (Gene Ontology (GO)) for data annotation and (iii) approaches that combine datamining and ontologies for functional analysis of gene expression data.

3. FUNCTIONAL GENE EXPRESSION DATA ANALYSIS BY DATAMINING

To answer the questions of biologists such as: are there clusters according to the genes expression profiles? What distinguishes these samples, these genes? Can we predict clusters, classifications? Datamining methods have been used to classify, aggregate and visualize these expression data.

Data mining is a process that is used to search through large amount of data in order to find useful information. Several data mining methodologies have been proposed to analyze large amounts of gene expression data. Most of these techniques can be broadly classified as cluster analysis and classification techniques. These techniques have been widely used to identify patterns expressions and co-expressed genes and to construct models able of predicting the behavior of genes. In this paper we focus on clustering, classification and association rule.

3.1. Clustering Techniques

Clustering has for objective to describe data independent of any a priori knowledge and to reduce the amount of data by categorizing or grouping similar data items together. To categories genes with similar functionality, various clustering methods are used:

- Hierarchical methods like agglomerative hierarchical clustering (AHC)
- Partitioned methods like K-means and C-fuzzy means,
- Model-based methods like self-organizing map (SOM)

Several works are considered to be the pioneers in this field [8, 9, 10]. Clustering was used on pharmacovigilance data [11] and in diagnosis of cancer [12]. Many comparative studies have been conducted to determine the most efficient clustering algorithm [13, 14, 15, and 16] but currently no consensus is established.

K-means method is used in various applications such as time-series yeast gene expression analysis [17] and the classification breast cancer subtypes [18]. However, in the real nature of biological data, a gene may be involved in several biological processes at once. Hence the use of the Fuzzy C-Means method [19] to give the possibility to a gene belonging to more than one expression profile at a time.

To conclude, firstly, we can say that clustering can work well when there is already a wealth of knowledge about the pathway in question, but it works less well when this knowledge is sparse [20]. And secondly several clustering algorithms have been proposed to analyze gene expression data. In general, there is no best clustering methods. They focus on models and characteristics of various data. Table 1, below shows a comparison of these techniques.

Table 1. Clustering methods comparison

Method	Principle	Example of use	Advantages	Desadvantages
K-Means	Decomposes the data set into a set of disjoint clusters: identifies subsets of genes with similar behavior.	Used by [21] to determine the expression profile during seven periods of the cell cycle in yeast.	Relatively efficient Easy implementation Allows to obtain a mean profile for each class Well suited to large data sets	Need to specify K, the number of clusters in advance Sensitive to noisy data and outliers Sensitive to start point Genes are forced to belong only to 1 cluster may not converge
Hierarchical clustering	Proceeds successively by either merging smaller clusters into larger ones, or by splitting larger clusters: it offers an intuitive visual distribution of the data	Used by [22] to classify and visualize dataset resulting from a proteomic analysis on species of pathogenic bacteria food-derived: *Listeria monocytogenes* and also used with the proteomics dataset to identify genes differentially expressed on sarcopenia in accordance with rat age	Does not require the number of clusters to be known in advance No input parameters (besides the choice of the similarity Computes a complete hierarchy of clusters	Not scale well: runtime No explicit clusters No automatic discovering of optimal clusters
Self organizing maps (SOM)	Partitioning experiments genes into a known number groups by association to nodes	Used by [23] to find groups of genes primarily involved in the differentiation mechanisms of enterocytes	The position of the groups space reflects the degree similarity between data. Data projected in a same neighborhood have close profiles expression. Insensitive to missing values	Need to specify the number of expected groups The results depend on the chosen distance
Fuzzy C-mean	Identifies genes pertaining to different regulatory clusters.	In [24], it provides a more interesting distribution of gene clusters compared to "ordinary" clustering methods when tested with melanoma and leukemia dataset.	Each gene can belong to multiple clusters.	No "natural" visualization of the data "Outlier" genes forced to belong to some cluster.

3.2. Classification Techniques

Classification employs a set of pre-classified data (training set) to develop a model that can classify the population of records at large. Among the most used methods are distinguished:

K-Nearest Neighbors (kNN): this method is very requested by biologists for its simplicity of interpretation. The classifier searches the k nearest neighbors of an unknown sample based on a distance measure. The most common metric used in Bioinformatics is the absolute Pearson

coefficient. For clinical end points and controls from breast cancer, neuroblastoma and multiple myeloma, authors in [25] generated 463,320 kNN models by varying feature ranking method, number of features, distance metric, number of neighbors, vote weighting and decision threshold. They identified factors that contribute to the MAQC-II project performance variation.

- **Support Vector Machines (SVM):** its principle is to search a hyper plane of optimal separation between two classes of sample space characteristics. This method was applied to the tumor classification from biochips. The SVM [26] or SVM combined with other techniques such as LDA [27] discriminate against non-linearly separable data and some of these approaches offer the possibility to define several classes. Other works have applied SVM with MI (Mutual Information) for the classification of colon cancer and Lymphoma [28]. But the disadvantage of this technique is to find the optimal separator border, from a set of learning in order to deal with cases where the data are not linearly separable. Another inconvenient, is the principle of a SVM is only applied to a problem with two classes. The generalization to multiple classes involves decomposition of the original problem into a set of sub binary problems between a particular class to the aggregation of all of the other classes ("one vs. all") or all classes "one versus one".

- **Decision trees (DT):** it is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables. The model built is in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Some authors, working on leukemia data (acute myeloid leukemia, acute lymphoblastic leukemia and chronic lymphoblastic), compared the performance of DT with the Subgroup Discovery Algorithms and SVM method [29]. According to the authors, DT gives good results. Other authors have combined a meta-heuristic called Particle Swarm Optimization (PSO) with DT (C4.5) and use it for patients' cancer data. They evaluated the performance of the proposed method (PSODT) and compare it with other algorithms of classification, such as: SOM, DT (C4.5), neural networks, SVM, and Naive Bayes. The results have shown that PSODT provides better than the others methods [30]

- **Association rule (AR):** An association rule is an expression of the form X □ Y, where X and Y are sets of items. Association is usually to find frequent item set findings among large data sets. Association Rule algorithm generate rules with confidence values. A study has been done in this regard by defining three different semantics addressing different biological goals: (1) similar expression levels between genes, (2) similar variations in expression levels of genes, (3) evolution in levels of gene expression. These rules have been applied to tumors breast and integrated in database software named MeV of the TIGR environment dedicated to the interpretation of microarray data [31]. The same authors made an improvement by adding rules for building regulatory networks from gene expression data filtered based on the five quality indices: support, confidence, lift, leverage and conviction [32]. .

3.3. Tools for Analysis of Gene Expression

There are an important range of tools for the application of classification methods and gene grouping. They include implementation of the main methods of clustering (hierarchical

clustering, k-means, SOM, etc.), accompanied by various graphical representations (heat maps, three-dimensional chart) facilitating the interpretation of the obtained results. In the table 2 we present examples of (software) tools for the classification and grouping of gene expression data.

Table 2. Tools/Environnement for gene classification and clustering

Software	**URL reference**
Weka	http://www.cs.waikato.ac.nz/ml/weka/
SAS Artificial	http://www.sas.com/technologies/analytics/datamining/miner/
IBM/SPSS Clementine	http://www.spss.com/software/modeling/modeler-pro/
SVMlight	http://svmlight.joachims.org
LIBSVM	http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/
Cluster and Treeview	http://rana.lbl.gov/ EisenSoftware.htm
	http://biosun1.harvard.edu/complab/dchip/
MeV	http://www.tm4.org/mev/.
MAGIC Tools	http://www.bio.davidson.edu/projects/magic/ magic.html

4. FUNCTIONAL GENE EXPRESSION DATA ANALYSIS BY THE USE OF ONTOLOGIES

The role of controlled vocabularies or ontologies is to capture the biological concepts describing biological objects such as genomic sequences, genes or gene products. These concepts are derived from publications of the results of the sequencing of genomes and their annotations. Therefore, the use of bio-ontologies becomes essential to deal with the heterogeneity of data and sources. It unifies the different definitions to improve the quality of data and facilitate the sharing and exchange of data.

4.1. Biological and Bioinformatic Ontologies

The Gene Ontology (GO) project [34] aims to provide a structured vocabulary to specific biological fields for describing gene products (protein or mRNA) function in the cellular context. It includes three parallel ontologies which are increasingly used by the bioinformatics community: (i) molecular functions, (ii) biological processes and (iii) cellular components. Terms are interconnected by relationship (is a, part of, regulates, etc).

GO is considered as the essential resource for the annotation. It is thus used by many portals (RefSeq, UniProt, KEGG, PDB, TAIR, etc.). Gene Ontology Annotation [35] is a portal dedicated to the data annotation of various interest organisms by using GO. AmiGO [36] is a portal that provides access to GO, it contains many cross-references with other information systems. The Open Biomedical Ontology project [37] is designed to create reference ontologies in biology and biomedical. The platform National Center for Biomedical Ontology (NCBO) [38] develops and maintains a web application called bioportal which allows researchers to access and use biomedical ontologies.

The Sequence Ontology (SO) project [39] was initially developed by the Gene Ontology Consortium for the definition of the characteristics of sequences that should be used in the annotation. It includes databases of model organisms such as WormBase, FlyBase, Mouse Genome Informatics group, and institutes such as the Sanger Institute and EBI.Other resources such as ArrayExpress at the EBI [40], GEO at NCBI [41], for the filing of data, expression of genes also contain information on the annotation of various organisms.

4.2. Ontologies of the Microarray Experiments

A formal description of experiences is extremely important for the organization and execution of experiments in biology. For example, the DNA chips for Micro-array Gene Expression Data project (MGED) [42] provide terms to annotate all aspects of an experience of DNA chips of its design with the definition of hybridization, to the preparation of the biological sample and the protocols used for hybridization on the chip and the analysis of data.

The terms MGED are organized in the form of ontology. It was built for the description of biological samples and their use in microarray experiments. This description focuses on biological material (biomaterials) and some treatments used during the experiment, thus, the ontology will be used directly by users to annotate their experiences on microarrays as well as developers of software and databases through structured queries experiences [43].

4.3. Semantic Similarity Measures

When biological entities are described using a common ontology, they can be compared by means of their annotations. This type of comparison is called semantic similarity. Several studies have been published describing and evaluating diverse semantic similarity measures. Semantic similarity has become a valuable tool for validating the results drawn from biomedical studies such as gene clustering, gene expression data analysis, prediction and validation of molecular interaction, etc.

The adoption of ontologies for annotation provides a means to compare entities on aspects that would otherwise not be comparable. For instance, if two gene products are annotated within the same schema, they can be compared by comparing the terms with which they are annotated [44]. The Gene Ontology is the main focus of investigation of semantic similarity in molecular biology because comparing gene products at the functional level is crucial for a variety of applications.

The authors in [44] give an interesting survey of semantic similarity measures applied to biomedical ontologies and describe examples of applications to biomedical research. As outlined by the authors, this survey will clarify how biomedical researchers can benefit from semantic

similarity measures and help them choose the approach most suitable for their studies. Several semantic similarity measures have been developed for use with GO. According to the strategies they employ, we distinguish:

4.3.1. Measures for comparing term

- **Node-based** [45, 46]: determines the information shared by two terms. A constraint of these measures is that they look only at a single common ancestor despite the fact that GO terms can have several disjoint common ancestors.

- **Edge-based** [47, 48, 49]: use the directed graph topology to compute distances between the terms to compare.

- **Hybrid** [50, 51]: combine different aspects of node-based and edge-based methods.

4.3.2. Measures for comparing gene products: to assess the functional similarity between gene products:

It is necessary to compare sets of terms rather than single terms. Several strategies have been proposed, they are grouped into two categories:

- **Pairwise** [44, 52]: measure functional similarity between two gene products by combining the semantic similarities between their terms.

- **Groupewise** [44, 53]: calculates directly similarity by one of three approaches: set, graph, or vector.

An early work was to measure the information content of the terms of the Gene Ontology (GO) [54]. Then it was evaluating some similarity measures such as Resnik, Lin and Jiang which are node-based measures on these annotated terms. Then, the same authors have investigated semantic similarity measures, and their application to ontological annotations of the SWISS-PROT database. They found a correlation between the semantic similarity of GO terms and the sequence similarity of the same genes aligned by BLAST [55].

In [56] controlled vocabularies containing medical concepts such as MeSH and SNOMED-CT were evaluated by a new measure based cross-modified path length feature between the concept nodes [56]. Afterwards, measures have been developed to take into account the fact that both terms can have several disjoint common ancestors (DCA) [57].

To overcome the weaknesses of the existing Gene Ontology browsers which use a conventional approach based on keyword matching, a genetic similarity measure is introduced in [58] to find a group of semantically similar Gene Ontology terms. The proposed approach combines semantic similarity measure with parallel genetic algorithm. The semantic similarity measure is used to compute the similitude strength between the Gene Ontology terms. Then, the parallel genetic algorithm is employed to perform batch retrieval and to accelerate the search in large search space of the Gene Ontology graph.

In [59] authors have attempted to improve existing measures such as the Wu Palmer measure by adding metadata by taking into account codes of evidence (codes that specify the quality of the annotation), the types of relationships between the GO terms deriving the metabolic pathways of different organisms (regulates, positively regulates, negatively regulates) and the qualifier NOT. This measure was applied to the metabolic pathways between species: human, mouse and the chicken [59].

However, although the Gene ontology, which is the reference for describing biological objects such as genome sequence, genes or gene products, it has only a static view of these biological objects and does not allow visualization that could express these concepts in space and time. Hence a combination of data mining to group similar expression profiles (static or temporal) and ontologies as additional annotation resources is desirable for the functional analysis of genes.

4. FUNCTIONAL GENE EXPRESSION DATA ANALYSIS BY DATAMINING AND BY THE USE OF ONTOLOGIES

Generally, the data analysis of expression takes place in two main steps: (1) identification of the groups of genes co-expressed, for example, by using clustering algorithms (2) functional analysis of these groups by using a controlled vocabulary such as the Gene Ontology (GO).

The following work [60] associates the first step to the second one. A transversal approach was developed based on the parallel grouping of the genes according to the biological annotations (vocabulary Gene Ontology), medical (UMLS terminology), genomic (characteristics of sequences) and experimental results (expression data).This approach has proved to be as powerful as a classical approach functioning in two phases. Others authors have suggested an approach based on fuzzy modelisation of differential expression profiles joined with data from GO, KEGG and Pfam [61]. An improvement of this approach was added by the same author by using the Formal Concepts Analysis method in upstream to get genes that have same expression profiles and same functional 'behaviour', and in downstream, it visualizes the results by Lattice [62].

5. DISCUSSION AND CONCLUSION

This article outlines various methods used in functional analysis of gene expression data.

At first, data mining methods, besides their diversity, appeared like a simple and obvious solution for determining expression profiles and the grouping or classification of genes with similar behavior. However, to ensure a complete analysis, we must give an annotation and a meaning to the results. That is to say bring semantics that could be achieved through controlled vocabularies such as GO and other sources of knowledge such as UniProt, KEGG, etc. Consequently, for better representation of co-expressed genes groups and a more "relevant" integration of genomic data supporting researchers in their experiments, recent works has been realized with both approaches.

As perspective, it would be interesting to do inter-species annotation on plants such as tomato because it contains a lot of anti-oxidants which protects from the ageing and certain cancers or on Medicago truncatula for its fixation of nitrogen in the soil with some model plants like Arabidopsis thaliana. The approach which will be used is the third one which employs data mining and ontologies for functional analysis of the expression data by accessing profile data of

expression and annotation via NCBI GEO, ArrayExpress sequence databases, using the Gene Ontology (GO) and Plant Ontology (PO) which includes terms on growth and stages of development of the plant and terms on the morphological and anatomical structures (tissues and cell types) of plants. The study will be on the aspect of space-time of terms by using Gene Ontology Annotation (GOA) as a resource.

REFERENCES

[1] Aharoni, A. & Vorst, O. (2002). "DNA microarrays for functional plant genomics". Plant Molecular Biology, Vol. 48, pp.99–118. DOI: http://dx.doi.org/10.1023/A:1013734019946

[2] Rensink, W. A. & Buell, C. R. (2005). "Microarray expression profiling resources for plant genomics". Trends in plant science, Vol. 10, no12, pp. 603-609. DOI: http://dx.doi.org/10.1016/j.tplants.2005.10.003.

[3] Grasso, L. C. Maindonald, J. Rudd, S, Hayward, D. C. Saint, R. Miller, & al., (2008). "Microarray analysis identifies candidate genes for key roles in coral development". BMC genomics, Vol. 9, no1, pp.1-18. DOI: http://dx.doi.org/10.1186/1471-2164-9-540.

[4] Guenin, L. Raharijaona, M. Houlgatte, R. & Baba-Aissa, F. (2010). "Expression profiling of prospero in the Drosophila larval chemosensory organ: Between growth and outgrowth". BMC genomics, Vol. 11, no1, pp.1-15, 2010. DOI: http://dx.doi.org/10.1186/1471-2164-11-47.

[5] Sharov, A.A .Piao,Y. Ko MS. "Gene expression profiling of mouse embryos with microarrays", Methods Enzymol, Vol. 477, pp. 511–541, 2010. DOI= https://dx.doi.org/10.1016/S0076-6879(10)77025-7.

[6] Govindarajan, R. Duraiyan, J. Kaliyappan, K. & Palanisamy, M. (2012). "Microarray and its applications", Journal of Pharmacy & Bioallied Sciences, 4(Suppl 2):S310-S312. DOI: http://doi.org/10.4103/0975-7406.100283.

[7] Médigue, C. Bocs, S. Labarre, L. Mathé, C. Vallenet, D. (2002) " The annotation in silico of genome sequences", MEDECINE/SCIENCES, Vol. 18, pp. 237-250.[original reference in French]

[8] Eisen, M. B. Spellman, P. T. Brown, P. O. & Botstein, D. (1998). "Cluster analysis and display of genome-wide expression patterns". Proceedings of the National Academy of Sciences, Vol. 95, no 25, pp. 14863-14868.

[9] Tamayo, P. Slonim, D, Mesirov, J, Zhu, Q. Kitareewan, S. Dmitrovsky, & al., (1999). "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation". Proceedings of the National Academy of Sciences, Vol.96, no 6, pp.2907-2912.

[10] Pan, W. Lin, J. & Le, C. T. (2002). "Model-based cluster analysis of microarray gene-expression data". Genome Biology, Vol. 3, Resarch0009. DOI: http://dx.doi.org/10.1186/gb-2002-3-2-research0009.

[11] Shannon, W. Culverhouse, R. & Duncan, J. (2003). "Analyzing microarray data using cluster analysis". Pharmacogenomics, Vol. 4, no1, pp.41–52. DOI: http://dx.doi.org/10.1517/phgs.4.1.41.22581.

[12] Smolkin, M. & Ghosh, D. (2003). "Cluster stability scores for microarray data in cancer studies". BMC bioinformatics, Vol.4, no1, pp.1-7. DOI: http://dx.doi.org/10.1186/1471-2105-4-36.

[13] Yeung, K. Y. Haynor, D. R. and Ruzzo, W. L. (2001). "Validating clustering for gene expression data".Bioinformatics, Vol. 17, pp. 309-318. DOI: http://dx.doi.org/10.1093/bioinformatics/17.4.309

[14] Dudoit, S. and Fridlyand, J. (2002), "A prediction-based resampling method for estimating the number of clusters in a dataset". Genome biology,Vol. 3, no 7, pp. RESEARCH0036-1-RESEARCH0036-21. DOI: http://dx.doi.org/10.1186/gb-2002-3-7-research0036.

[15] Romualdi, C. Campanaro, S. Campagna, D. Celegato, B. Cannata, N. Toppo, S. Valle, G. Lanfranchi, G. (2003). "Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification". Human Molecular Genetics, Vol. 12, no 8, pp. 823-836. DOI: http://dx.doi.org/10.1093/hmg/ddg093.

[16] Priness, I. Maimon, O. and Ben-Gal, I. (2007). "Evaluation of gene-expression clustering via mutual information distance measure", BMC Bioinformatics, Vol. 8, no 1, pp. 1-12, DOI: http://dx.doi.org/10.1186/1471-2105-8-111.

[17] Tavazoie, S. Hughes, J. D. Campbell, M. J. Cho, R. J.and Church, G. M. (1999). "Systematic determination of genetic network architecture". Nature genetics, Vol. 22, no 3, pp. 281-285. DOI: http://dx.doi.org/10.1038/10343M3

[18] Masuda, H. Baggerly, K. A.Wang, Y. Zhang, Y, Gonzalez-Angulo, and al., (2013). "Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes". Clinical Cancer Research, Vol 19, no19, pp.5533-5540. DOI:http://dx.doi.org/10.1158/1078-0432.CCR-13-0799

[19] Maji, P. and Paul, S. "Clustering Rough Sets Fuzzy Sets Microarray", (2012). Book "Perception and Machine Intelligence", Vol. 7143, pp. 203-210, Springer, 2012.

[20] Fiehn, O., Kloska, S., & Altmann, T. (2001). "Integrated studies on plant biology using multiparallel techniques". Current Opinion in Biotechnology, Vol 12, no1, pp.82-86. DOI: http://dx.doi.org/10.1016/S0958-1669(00)00165-8.

[21] Anusuya, S. Bhanu, D. N. U. and Kasthuri, E. (2015). "yeast gene expression analysis using k means and FCM", International Journal of Pharma and Bio Sciences, Vol. 6, no.3: B, pp. 395 – 400.

[22] Meunier, B. Dumas, E. Piec, I. Bechet, D. Hebraud, M. and Hocquette, J. F. (2007). "Assessment of Hierarchical Clustering Methodologies for Proteomic Data Mining", Journal of Proteome Research, Vol. 6, pp. 358-366. DOI: http://dx.doi.org/10.1021/pr060343h.

[23] H. Bedrine-Ferran, N. Le Meur, I. Gicquel, M. Le Cunff, N. Soriano, I. Guisle, and al., (2004). "Transcriptome variations in human CaCo-2 cells: a model for enterocyte differentiation and its link to iron absorption", Genomics, Vol. 83, no 5, pp. 772-789.
DOI: http://dx.doi.org/10.1016/j.ygeno.2003.11.014.

[24] Seo Young, K. and Tai Myong, C. (2007). "Fuzzy Types Clustering for Microarray Data", International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering, Vol. 1, no. 4, pp. 229-232, 2007.

[25] Parry, R. M. Jones, W. Stokes, T. H. Phan, J. H. Moffitt, R. A. Fang, H. Shi, L. Oberthuer, A. Fischer, M. Tong, W. Wang, M. D. (2010). "K-Nearest Neighbor Models for Microarray Gene Expression Analysis And Clinical Outcome Prediction", Pharmacogenomics Journal, Vol. 10. no.4, pp. 292–309. DOI: http://dx.doi.org/10.1038/tpj.2010.56.

[26] Li, F. and Yang, Y. (2005). "Analysis of recursive gene selection approaches from microarray data". Bioinformatics, Vol. 21, no.19, pp. 3741-3747.DOI: http://dx.doi.org/10.1093/bioinformatics/bti618.

[27] Niijima, S. and Kuhara, S. (2006). "Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE", Biomedcentral, Vol.7. pp. 1-18. DOI: http://dx.doi.org/ doi: 10.1186/1471-2105-7-543

[28] Vanitha, C. D. A. Devaraj, D. and Venkatesulu, M. 2015. Procedia Computer Science, (2015), "Gene Expression Data Classification using Support Vector Machine and Mutual Information-based Gene Selection", Procedia Computer Science, Vol. 47, pp. 13-21.
DOI: http://dx.doi.org /doi:10.1016/j.procs.2015.03.178.

[29] Netto, O. P. Nozawa, S. R. Mitrowsky, R. A. R. Macedo, A. A. and Baranauskas, J. A. (2010). "Applying decision trees to gene expression data from dna microarrays: A leukemia case study". In XXX Congress of the Brazilian Computer Society, X Workshop on Medical Informatics, pp.1-10.

[30] Chen, K. H. Wang, K. J. Tsai, M. L. Wang, K. M. Adrian, A. M., Cheng, and al. (2014). "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm". BMC bioinformatics, Vol. 15, no 49, 1-10. DOI: http://dx.doi.org/10.1186/1471-2105-15-49.

[31] Agier, M. Petit, J-M. Chabaud, V. Pradeyrol, C. Y-Bignon and Vidal, V. (2004). "Different types of rules for expression of genes database Application to database of mammaire tumor", In XXIIème Congrès INFORSID, pp. 351–367. Biarritz : France. [original reference in French].

[32] Agier, M, "Different types of rules for the reconstruction of networks of genes from expression data". (2007). Revue I3 Information Interaction-Intelligence, numéro hors série, pp. 161-81, Cépaduès Editions. [original reference in French].

[33] Selvaraj, S. and Natarajan, J. (2011). "Microarray Data Analysis and Mining Tools", Biomedical Informatics, Vol. 6, no 3, pp. 95-99. DOI: http://dx.doi.org/10.6026/97320630006095.

[34] T.Z. Berardini, T.Z. Li, D. Huala, E. Bridges, S. Burgess, S. McCarthy, F. and al. 2010. "The Gene Ontology in 2010: extensions and refinements", Nucleic Acids Res, Vol. 38, (Database issue): D331-D335. (cf: http://www.geneontology.org).

[35] Huntley, R. P. Sawford, T. Mutowo-Meullenet, P. Shypitsyna, A. Bonilla, C. Martin, M. J. & O'Donovan, C. (2015). The GOA database: gene ontology annotation updates for 2015. Nucleic acids research, Vol. 43(D1), pp. D1057-D1063. (cf: http://www.ebi.ac.uk/GOA).

[36] Carbon, S. Ireland, A. Mungall, C. J. Shu, S. Marshall, B. Lewis, S., & Web Presence Working Group. (2009). AmiGO: online access to ontology and annotation data. Bioinformatics,Vol. 25, no. 2. pp. 288–289. DOI: http://dx.doi.org/10.1093/bioinformatics/btn615

[37] Ghazvinian, A. Noy, N. F. & Musen, M. A. (2011). How orthogonal are the OBO Foundry ontologies? Journal of biomedical semantics, Vol.2, (Suppl2):S2.
DOI: http://dx.doi.org/10.1186/2041-1480-2-S2-S2.

[38] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and al. (2011). BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research, Vol. 39 (Web Server issue):W541-W545.

[39] Eilbeck, K. Lewis, S.E. Mungall, C.J. Yandell, M Stein, L. Durbin, R. Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. Genome Biology, Vol. 6, no. 5: r44. DOI: http://dx.doi.org/10.1186/gb-2005-6-5-r44.

[40] Parkinson, H. Kapushesky, M. Shojatalab, M. Abeygunawardena, N. Coulson, R. Farne, A. E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, A. Brazma. (2007). ArrayExpress—a public database of microarray experiments and gene expression profiles. Nucleic acids research, Vol. 35(suppl 1), pp. D747-D750. DOI: http://dx.doi.org/10.1093/nar/gkl995.

[41] Barrett, T.Troup, D. B. Wilhite, S. E. Ledoux, P. Rudnev, D. Evangelista, and Edgar, R. (2007). "NCBI GEO: mining tens of millions of expression profiles—database and tools update", Nucleic Acids Research, Vol. 35, pp. D760- D765. DOI: http://dx.doi.org/10.1093/nar/gkl887.

[42] Guérin, E., Marquet, G., Burgun, A., Loréal, O., Berti-Equille, L., Leser, U., & Moussouni, F. (July 2005). "Integrating and warehousing liver gene expression data and related biomedical resources in GEDAW". In International Workshop on Data Integration in the Life Sciences, pp. 158-174. Springer Berlin Heidelberg.

[43] Griffin, J. L, & Steinbeck, C. (2010). "So what have data standards ever done for us? The view from metabolomics". Genome Medicine,Vol. 2, no.6:38, pp.1-3. DOI: http://dx.doi.org/10.1186/gm159.

[44] Pesquita, C. Faria, D. Falcao, A. O. Lord, P. & Couto, F. M. (2009). "Semantic similarity in biomedical ontologies". PLoS Comput Biol, Vol. 5, no.7: e1000443. DOI=http://dx.doi.org/ 10.1371/journal.pcbi.1000443.

[45] Resnik, P. (1999). "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language". Journal of Artificial Intelligence 11, pp.95–130.

[46] Lin, D. (July 1998), "An Information-Theoretic Definition of similarity", In: Proceedings of The Fifteenth International Conference on Machine Learning (ICML'98), pp. 296-304.

[47] Rada, R. Mili, H. Bicknell, E. & Blettner, M. (1989). "Development and application of ametric on semantic nets", IEEE Transaction on Systems, Man, and Cybernetics, Vol 19, no. 1, pp.17–30.

[48] Wu, Z., & Palmer, M. (June 1994). "Verb semantics and lexical selection". In: Proceedings of The 32nd Annual Meeting of the Associations for Computational Linguistics, 1994, pp. 133–138.

[49] Hirst, G., & Budanitsky, A. (2005). "Correcting real-word spelling errors by restoring lexical cohesion. Natural Language Engineering, Vol. 1, no. 1, pp 87-111.

[50] Jiang, J. J. & Conrath, D. W. (1997). "Semantic similarity based on corpus statistics and lexical taxonomy", In: Proceedings of The International Conference on Research in Computational Linguistics, arXiv preprint cmp-lg/9709008. Taiwan.

[51] Leacock, C., & Chodorow, M. (1998). "Combining Local Context and WordNet Similarity for Word Sense Identification. In WordNet: An Electronic Lexical Database", Vol.49, no. 2, pp. 265-283.

[52] Chagoyen, M., Carazo, J., & Pascual-Montano, A. (2008). "Pairwise similarity scores using functional annotations: review and comparison". In 8th Spanish Symposium on Bioinformatics and Computational Biology: 2008.

[53] Teng, Z. Guo, M. Liu, X. Dai, Q. Wang, C. & Xuan, P. (2013). "Measuring gene functional similarity based on group-wise comparison of GO terms". Bioinformatics, pp. 1-9. DOI: http://dx.doi.org/10.1093/bioinformatics/btt160.

[54] Lord, P. W. Stevens, R. D. Brass, A. & Goble, C. A. (October 2003). "Semantic similarity measures as tools for exploring the gene ontology". In Pacific Symposium on Biocomputing, Vol. 8, no. 4, pp. 601-612.

[55] Lord, P. W. Stevens, R. D. Brass, A. & Goble, C. A. (2003). "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation". Bioinformatics, Vol. 19, no. 10, pp.1275-1283. DOI: http://dx.doi.org/10.1093/bioinformatics/btg153.

[56] H. Al-Mubaid, and H.A. Nguyen. (August 2006). "A cluster-based approach for semantic similarity in the biomedical domain", In Engineering in Medicine and Biology Society, 2006. EMBS'06. 28th Annual International Conference of the IEEE, pp. 2713-2717.

[57] Couto, F. M., Silva, M. J., & Coutinho, P. M. (October 2005). "Semantic similarity over the gene ontology: family correlation and selecting disjunctive ancestors", In Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 343-344. DOI: http://dx.doi.org/10.1145/1099554.1099658.

[58] Othman, R. M. Deris, S. & Illias, R. M. 2008. "A genetic similarity algorithm for searching the Gene Ontology terms and annotating anonymous protein sequences", Journal of biomedical informatics, Vol. 41, no.1, pp. 65-81. DOI: http://dx.doi.org/10.1016/j.jbi.2007.05.010.

[59] Bettembourg, C. Diot, C. & Dameron, O. (2014). "Semantic particularity measure for functional characterization of gene sets using gene ontology". PloS one, Vol. 9, no.1, e86525. DOI: http://dx.doi.org/10.1371/journal.pone.0086525.

[60] Chabalier, J. Mosser, J. & Burgun, A. (2007). "A transversal approach to predict gene product networks from ontology-based similarity". BMC bioinformatics, Vol. 8, no. 235, pp 1-12. DOI: http://dx.doi.org/10.1186/1471-2105-8-235.

[61] Devignes, M. D. Benabderrahmane, S. Smaïl-Tabbone, M. Napoli, A. & Poch, O. (2012). "Functional classification of genes using semantic distance and fuzzy clustering approach: evaluation with reference sets and overlap analysis". International journal of computational biology and drug design, Vol. 5, no. 3-4, pp. 245-260. DOI: http://dx.doi.org/10.1504/IJCBDD.2012.049207.

[62] Benabderrahmane, S. "Formal Concept Analysis and Knowledge Integration for Highlighting Statistically Enriched Functions from Microarrays Data". (2014). International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2014, Granada, Spain, Granada.

RECOGNITION OF RECAPTURED IMAGES USING PHYSICAL BASED FEATURES

S. A. A. H. Samaraweera[1] and B. Mayurathan[2]

[1]Department of Computer Science, University of Jaffna, Sri Lanka
anuash119@gmail.com
[2]Department of Computer Science, University of Jaffna, Sri Lanka
barathy@jfn.ac.lk

ABSTRACT

With the development of multimedia technology and digital devices, it is very simple and easier to recapture a high quality images from LCD screens. In authentication, the use of such recaptured images can be very dangerous. So, it is very important to recognize the recaptured images in order to increase authenticity. Image recapture detection (IRD) is to distinguish real-scene images from the recaptured ones. An image recapture detection method based on set of physical based features is proposed in this paper, which uses combination of low-level features including texture, HSV colour and blurriness. Twenty six dimensions of features are extracted to train a support vector machine classifier with linear kernel. The experimental results show that the proposed method is efficient with good recognition rate of distinguishing real scene images from the recaptured ones. The proposed method also possesses low dimensional features compared to the state-of-the-art recaptured methods.

KEYWORDS

Image Recapture Detection, Texture, HSV, Blurriness & Support Vector Machine

1. INTRODUCTION

Since the last century, the information technology is increasing rapidly. The digital documents are replacing paper documents. However, a photograph implies truthfulness. This technology enables digital documents to be easily modified and converted which makes our life easier in digital matters. Unlike a text, an image accomplishes an effective communication channel for humans. Hence, maintenance of the trustfulness of a digital image is a major challenge in today's world. Recaptured images means, it is different from the common photographs in that what being captured is an image reproduction surface instead of a general scene. Image recapture detection technique distinguishes real images from recaptured images. i.e.) images from media that displays real-scene images such as printed pictures or LCD display. Difficulties of recognizing recaptured images can be described using Figure 1. Here (a) and (b) are real images, (c) and (d) are recaptured images. It is extremely complicated task for an artificial system to recognize recaptured images from real ones.

Jan Zizka et al. (Eds) : ACSTY, NATP - 2016
pp. 15– 31, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61402

In recent years, considerable amount of researches are conducting for image recaptured detection to restore the trustworthiness of digital images [1], [2], [3]. Using the image recapturing process it is possible to restore the intrinsic image regularities automatically and to remove some common tampering anomalies automatically. An important task for the current image forensic system is the recognition of the recaptured images. Apart from that, an image forensic system can detect rebroadcast attacks on a biometric identification system. Therefore we study the problem of recaptured image detection as an application in image forensics.

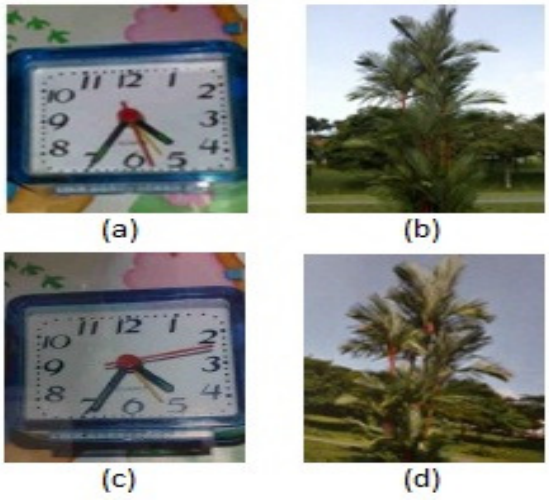

Figure 1. Difficulties of recognizing recaptured images

In other hand, face authentication systems are designed with aliveness detection for verifying a live face on mobile devices such as laptop computers and smart phones. For such systems faked identity through recapturing of a faked print face photo has become a big issue.

In robot vision, differentiating the objects on a poster from the real ones is more intelligence. IRD is also useful for that purpose. Another important application for IRD is in composite image detection. One way to cover composition in composite image is to recapture it.

The process of producing the real scene images and the corresponding recaptured images are shown in Figure 2. As shown in Figure 2. (a) the real image can be obtained through any type of camera. For the reproduction process, initially the real image is captured by the any type of camera. Then it is reproduced using different types of printing or display media such as printed on an office A4 paper using a colour laser printer or displayed on a LCD screen of a PC etc. Finally, the recaptured image is obtained through the camera.

Displaying or printing a scene on any type of physical media, lead to poor quality of recaptured image. We can easily identify some artefacts like texture pattern, colour fading etc. As shown in Figure 3.the low-quality recaptured images can be easily identified by the human eyes.

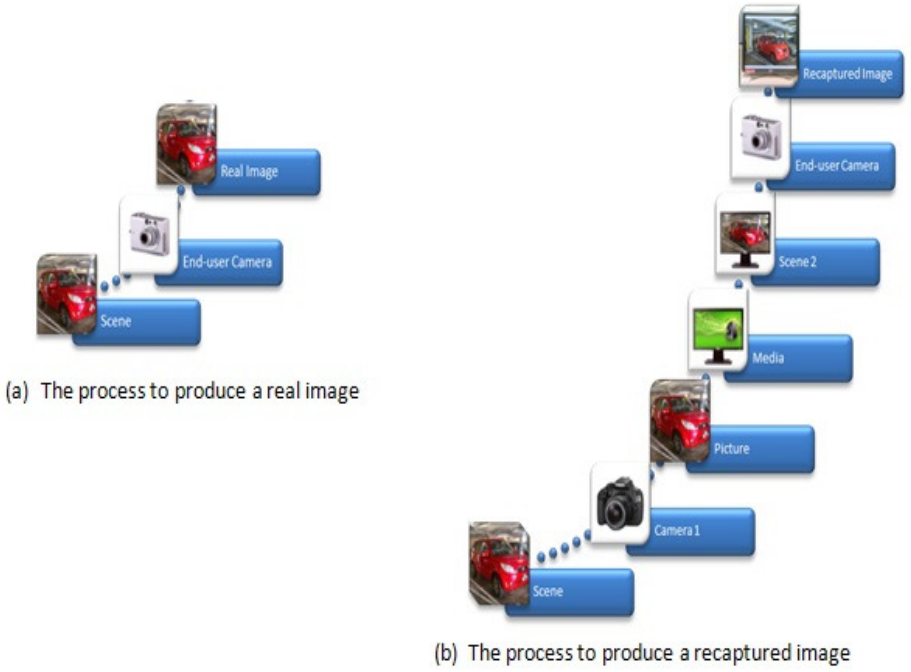

Figure 2. The process of producing real image and recaptured image

Figure 3. Comparison of a real image (a) and a recaptured image (b)

For instance, consider the displaying on LCD screen as the reproduction process as illustrated in Figure 4. Cao and Kot [5] compared real images and corresponding recaptured images with a large number of controllable settings including camera settings, LCD settings and environmental settings. They concluded that visual quality of these finely recaptured images is significantly better than the casually recaptured images. So, this is a big opportunity for forgers to recapture the artificially generated scenery and use the recaptured image to fool image forensic system. Recently, a Vietnamese security group found that most commercial laptop computers with face authentication system can be easily attacked by just presenting a human face printed on an A4-size paper [6].

Figure 4. Some controllable settings for reproduction process on a LCD screen

2. LITERATURE REVIEW

This section includes several approaches which are used to identify the recaptured images from real scene images as well as the studies which are related with distinguishing real scene images and recaptured images on the printing paper and LCD screens respectively.

Xinting Gao et al., [1] introduced a physics-based approach for recaptured image detection. The set of physics-based features is composed of the contextual background information, the spatial distribution of specularity that is related to the surface geometry, the image gradient that captures the non-linearity in the recaptured image rendering process, the colour information and contrast that is related to quality of reproduction rendering, and a blurriness measure that is related to the recapturing process. These features were used to classify the recaptured images from the real ones. This achieved significantly better classification performance on the low resolution images as compared to the wavelet statistical features.

Ke et al., [2] proposed an image recapture detection method based on multiple feature descriptors. It uses combinations of low dimensional features including texture feature, noise feature, difference of histogram feature and colour feature. The experimental result has demonstrated that this method is efficient with good detection rate of distinguishing real scene images from the recaptured ones. It possesses low time complexity.

Hany Faridy and Siwei Lyu [4] presented a statistical model with first and higher order statistics which capture certain statistical regularities of natural images.

Hang Yu et al., [3] brought up a cascaded dichromatic model with the high frequency spatial variations in the specular component of a recaptured image. This distinctive feature is a result of the micro-structure on the printing paper. With a probabilistic support vector machine classifier, Cao and Kot [5] classified recaptured images on LCD screens from natural images. They perform the experiment using three types of features including texture feature using Local Binary Pattern, multi-scale wavelet statistics and colour feature.

3. DATASET

Bai et al., [7] found that the image resolution affects the performance of the algorithms. So, XintingGao et al., [8] presented smart phone recapture image database taken by smart phone cameras. Even though there are some publically available databases, I used this database due to the general resolution of the images is set to VGA (640 x 480). This database has constructed using following criteria.

- The images are in pair for the real image and the recaptured one taken by the same end-user camera.
- The images are consisting of outdoor natural scene, indoor office or residence scene and close-up or distant scene

3.1. Real Image Dataset

The real images are obtained by any type of camera as shown in Figure 1 (a). The images in the real image dataset have produced using three popular brands of smart phones including Acer M900, Nokia N95 and HP iPAQ hw6960. These camera phones are set to auto mode whenever possible. All these three types of phones have back-facing camera. Totally I used 1094 images as real images. Table 1 lists total number of images taken from different brands of camera.

Table 1. The number of real images.

Types	Images
Acer B	407
HP B	369
Nokia B	318
Total	1094

3.2. Recaptured Image Dataset

As illustrated in Figure 1 (b), the reproduction process is pure image-based. The images in the recaptured image dataset have produced using three types of DSLR (digital single-lens reflex) cameras including Nikon D90, Canon EOS 450D and Olympus E-520. These cameras are set to auto mode whenever possible and the resulting images have saved in JPEG format. The DSLR cameras have high resolution (greater than 3000x2000 pixels) and high quality. In constructing the recaptured dataset it has used two types of reproduction processes such as printing on a paper and displaying on a screen. The images are printed on an A4-size office paper using HP CP3505dn laser printer and Xerox Phaser 8400 ink printer. They have printed into 4R glossy and matte photos too. On the other hand, for LCD screen display they have used Dell 2007FP LCD screen (1600 x 1200 pixels). Finally the reproduced image has recaptured by the above mentioned camera phones. Table 2 lists the number of recaptured images in each reproduction process. Totally I used 1137 images as recaptured images.

Table 2. The number of recaptured images.

Types of camera phones	Types	Images	Total
Acer B	LCD – NikonSLR	9	269
	PhotoGlossy – NikonSLR	20	
	PhotoGlossy - OlympusSLR	30	
	PhotoMatte - NikonSLR	30	
	PhotoMatte - OlympusSLR	37	
	PrintInk - OlympusSLR	50	
	PrintLaser - NikonSLR	93	
HP B	LCD – NikonSLR	06	398
	PhotoGlossy – CannonSLR	76	
	PhotoMatte - CannonSLR	73	
	PrintInk - CannonSLR	65	
	PrintLaser - CannonSLR	130	
	PrintLaser - NikonSLR	48	
Nokia B	PhotoGlossy – OlympusSLR	46	470
	PhotoMatte – OlympusSLR	80	
	PrintInk - OlympusSLR	169	
	PrintLaser - NikonSLR	35	
	PrintLaser - OlympusSLR	140	
Scenery in Total			1137

4. METHODOLOGY

In this paper, I propose an image recaptured detection method based on physical based features. A working diagram of my proposed method is illustrated in Figure 5. The images in the real image dataset and the recaptured image dataset are used for the Feature Extraction step. For each image, features including Texture, HSV colour and Blurriness are extracted. Then to train the SVM classifier, both features and labels are used. This is the training procedure in my method. In the testing procedure, the features in the testing image are extracted. Then the SVM classifier classifies those features as the features of either a real image or a recaptured image.

4.1. Feature Extraction

In general, the recaptured images and corresponding real images will never be same due to the direction of the light, distance between the camera and the scenery, sensor resolution, the lens quality and so forth. By considering this problem as a binary classification task, I introduce following three types of features including Texture, HSV colour and Blurriness to differentiate the recaptured images from real images.

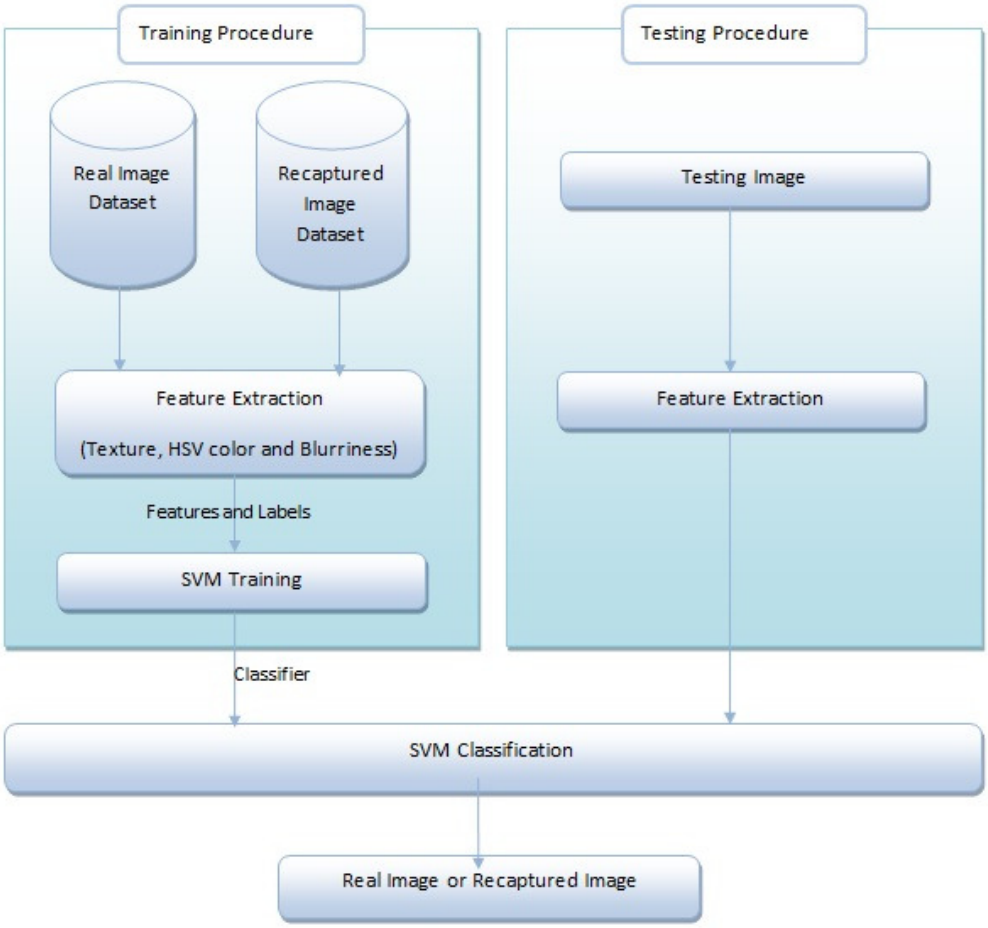

Figure 5. Diagram for the proposed image recaptured detection method

4.1.1. Texture feature

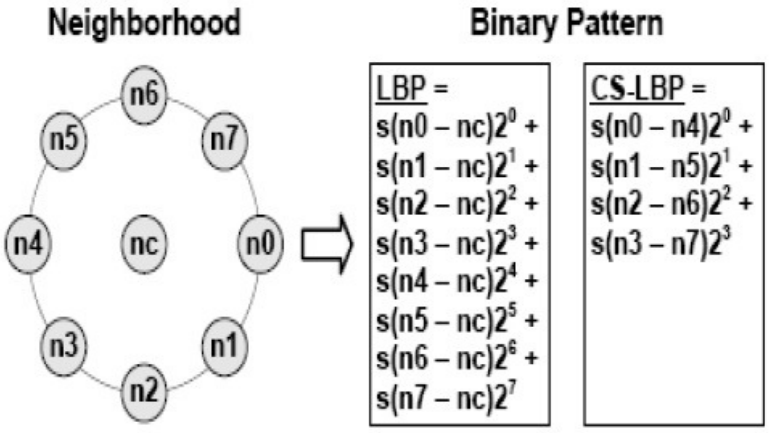

Figure 6. LBP and CS-LBP features for a neighbourhood of 8 pixels

In computer vision applications, Texture plays an important role. During the past decades, so many numbers of algorithms have been presented for texture feature extraction. They can be mainly divided into two approaches: Statistical approaches and Structural approaches. Among them most commonly used algorithms are Gabor filters, Wavelet transform and so forth. Currently the local binary pattern (LBP) has received a considerable attention in many applications as a Statistical approach [9]. Due to the high dimensionality of the LBP operator, now new experiments are carrying on with the centre-symmetric local binary pattern (CS-LBP) which is an extension of LBP operator. Not only dimensionality reduction, the CS-LBP captures better the gradient information than the basic LBP. Since the CS-LBP descriptor is computationally simple, effective and robust for various image transformations, it is very important to present a brief review of the CS-LBP.

CS-LBP operator [10] initially leads by the LBP operator. Histograms of the LBP operator are long (256) and it is not robust in flat images. CS-LBP was proposed to reduce these drawbacks.

The LBP operator compares each pixel with the centre pixel. Instead of that, the CS-LBP operator compares centre-symmetric pairs of pixels as illustrated in Figure 6. For the same number of neighbours, it produces half number of comparisons. So that the LBP produces 256 (2^8) different binary patterns, whereas the CS-LBP produces only 16 (2^4) different pattern for 8 neighbours. For flat areas, the operator's robustness can be increased using the gray level differences that are threshold at a small value T. Thus, the CS-LBP operator is defined by Eq. (1).

$$CS-LBP_{R,N,T}(x,y) = \sum_{i=0}^{\frac{N}{2}-1} s(n_i - n_{i+(N/2)}) \times 2^i, s(x) = \begin{cases} 1 & \text{if } x \succ T \\ 0 & \text{Otherwise} \end{cases} \quad \text{Eq.(1)}$$

Where

(x, y) denotes the coordinates of a pixel,

n_i and $n_{i+\frac{N}{2}}$ corresponds to the gray level of the center-symmetric pairs of pixels of N equally spaced pixels on a circle of radius R.

4.1.2. Colour feature

In the reproducing stage, the reproduction devices introduce some tint into the reproduced images. And also, the lighting can reduce the contrast and saturation of a recaptured image. So, the colour feature of a recaptured image looks different from its original image as shown in Figure 7.

Figure 7. Comparison of the colour features introduced by the reproduction process

Colour model describes colours. Usually colour models represent a colour in the form of tuples (generally of three). The purpose of a colour model is to facilitate the specification of colours in a certain way and common standard. The RGB colour model is the most common colour model for digital images. Because it retains compatibility with computer displays. However RGB has some drawbacks. RGB is non-useful for objects specification and recognition of colours. It is difficult to determine specific colour in RGB model. It reflects the use of CRTs, since it is hardware oriented system. Apart from RGB the HSV colour model is commonly used in colour image retrieval system, since HSV colours are defined easily by human perception not like RGB.

The HSV stands for the Hue, Saturation, and Value. The coordinate system is in a hexagon in Figure 8. (a). And Figure 8. (b) shows a view of the HSV colour model. The Value represents intensity of a colour, which is decoupled from the colour information in the represented image. The hue and saturation components are intimately related to the way human eye perceives colour resulting in image processing algorithms with physiological basis. As hue varies from 0 to 1.0, the corresponding colours vary from red, through yellow, green, cyan, blue, and magenta, back to red, so that there are actually red values both at 0 and 1.0. As saturation varies from 0 to 1.0, the corresponding colours (hues) vary from unsaturated (shades of gray) to fully saturated (no white component). As value, or brightness, varies from 0 to 1.0, the corresponding colours become increasingly brighter.

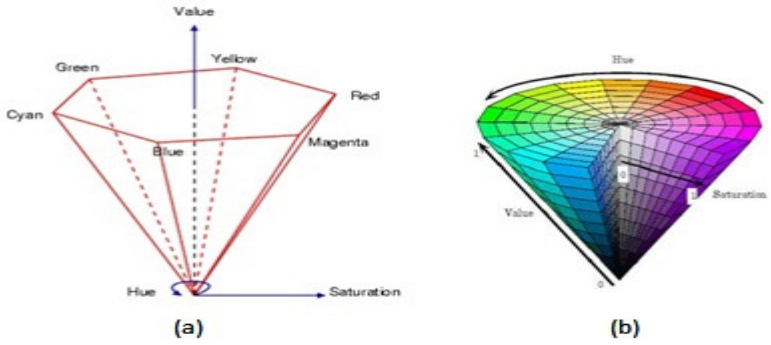

Figure 8. (a) HSV Cartesian Coordinate System (b) HSV colour model

Colour histogram and colour moments are widely used to represent the colour information of an image. Colour histogram is the approach more frequently adopted for Content Based Image Retrieval Systems. It describes the frequency of colours in images. Even though it is a widely used feature, it has some disadvantages associated with it. It is sensitive to noisy interferences.

Small change in image might result in large change in histogram values and it is computationally expensive.

Colour moments are measures that can be used to differentiate images based on their features of colour. The assumption of the basis of colour moments is that the distribution of colour in an image can be interpreted as a probability distribution. Probability distributions are characterized by a number of unique moments. For example normal distributions are differentiated by their mean and variance. Therefore it follows that if the colour in an image follows a certain probability distribution, the moments of that distribution can then be used as features to identify that image based on colour.

The mean, standard deviation and Skewness of an image are known as colour moments. In HSV colour model, a colour is defined by 3 values: Hue, Saturation, and Value. Colour moments are calculated for each of these channels in an image. An image therefore is characterized by 9 moments: 3 moments for each 3 colour channels. We will define the i^{th} colour channel at the j^{th} image pixel as p_{ij}. The three colour moments can be defined as:

- Moment 1- Mean:

$$E_i = \sum_{j=1}^{N} \frac{1}{N} p_{ij} \qquad \text{Eq.(2)}$$

  Mean can be described as the average colour value in the image

- Moment 2- Standard Deviation:

$$\sigma_i = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^2} \qquad \text{Eq.(3)}$$

  The standard deviation is the square root of the variance of the distribution

- Moment 3- Skewness:

$$s_i = \sqrt[3]{\frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^3} \qquad \text{Eq.(4)}$$

Skewness can be described as a measure of the degree of asymmetry in the distribution.

For example in HSV colour space, the variable i can take values from 1 to 3 (i.e. 1=H, 2=S, 3=V). So, the resultant feature for the image contains 9 values in the form of 3x3 matrix of the following format:

$$\begin{bmatrix} E_{11} & E_{12} & E_{13} \\ \sigma_{11} & \sigma_{12} & \sigma_{13} \\ s_{11} & s_{11} & s_{13} \end{bmatrix}$$

Where:

E_{11},E_{12},E_{13} represents Mean value for HSV.

$\sigma_{11},\sigma_{12},\sigma_{13}$ represents Standard deviation value for HSV.

s_{11},s_{12},s_{13} represents Skewness value for HSV.

4.1.3. Blurriness

In a recaptured image, there are three key factors that blurriness can arise:

- The first capture device or the printing device could be of low resolution.
- The display medium may not be in the focus range of the camera due to specific recaptured settings.
- If the end user camera has a limited depth of field, the distance background may be blurring, while the entire display medium is in focus.

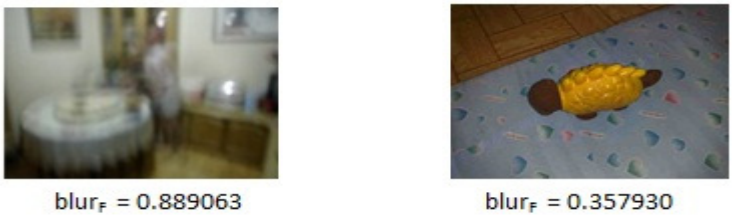

Figure 9. (a) and (b)

In this research work, I explore such information as a distinguishing feature in order to recognize whether an image is a real scene or recaptured image. The new method based on the discrimination between different levels of blur perceptible on the same image proposed by Crete et al. [11] have been used to calculate a no-reference perceptual blur metric, ranging from 0 to 1 which are respectively the best and the worst quality in term of blur perception as shown in Figure 9. (a) and (b) respectively.

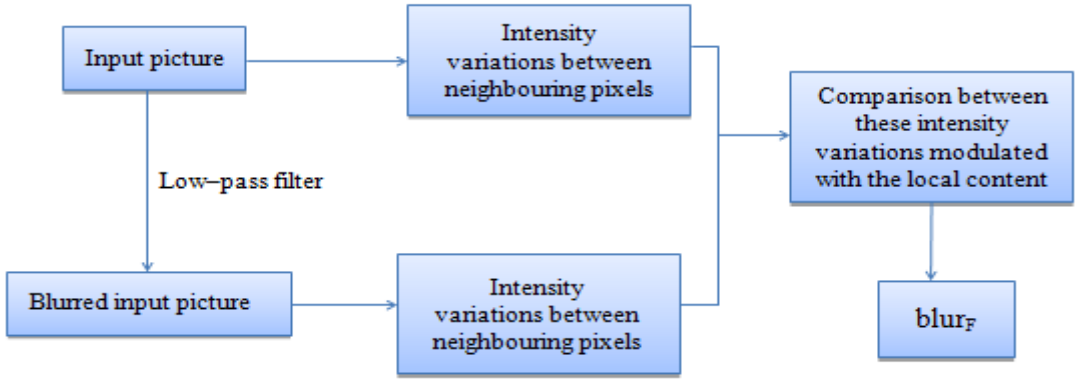

Figure 10. Simplified flow chart of the blur estimation principle

As shown in Figure 10in the first step the intensity variations between neighbouring pixels of the input image is computed. Then a low-pass filter is applied and computed the variations between the neighbouring pixels. Then, the comparison between these intensity variations allows us to

evaluate the blur annoyance. Thus, a high variation between the original and the blurred image means that the original image is sharp whereas a slight variation between the original and the blurred image means that the original image is already blurred.

4.2. Classification

Classification consists of predicting a certain outcome based on a given input. Among various classification techniques, Support Vector Machine (SVM) is originally developed for solving binary classification problems [12].

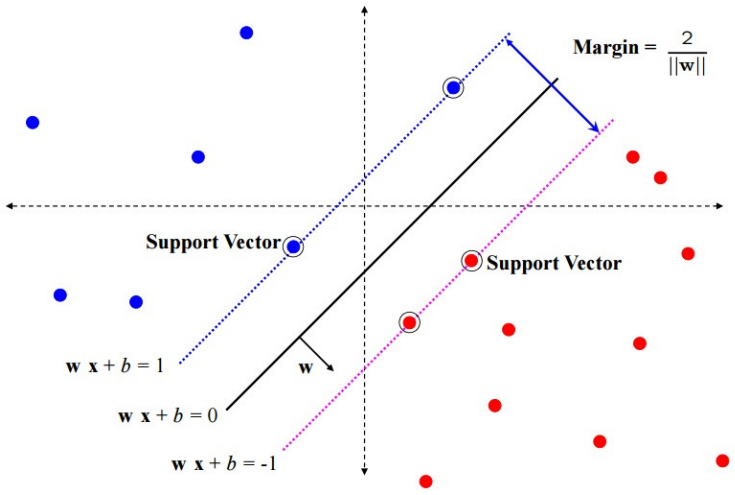

Figure 11. An example of a separable problem in a 2 dimensional space

Consider the Figure 11as an example of a linearly separable problem. Suppose we are given a set of l training points of the form:

$$(x_1, y_1), (x_2,y_2), \ldots, (x_i ,y_i), \ldots, (x_l ,y_l) \in \Re^n \times \pm 1$$

Where

x_i is an n-dimensional vector and

y_i are their labels such that:

$$y_i = \begin{cases} +1 & ; \text{if the vector is classified to class } +1 \\ -1 & ; \text{if the vector is classified to class -1} \end{cases}$$

We thus try to find a classification boundary function f(x) = y that not only correctly classifies the input patterns in the training data but also correctly classifies the unseen patterns. The classification boundary f (x) = 0, is a hyperplane defined by its normal vector w, which basically divides the input space into the class +1 vectors on one side and the class -1 vectors on other side. Then there exists f (x) such that

$$f(x) = w.x + b, w \in \Re^n \text{ and b } \in \Re \qquad \text{Eq.(5)}$$

subject to

$$y_i f(x_i) \geq 1 \text{ for i} = 1, 2, \ldots, \text{n} \qquad \text{Eq.(6)}$$

The optimal hyperplane is defined by maximizing the distance between the hyperplane and the data points closest to the hyperplane (called support vectors). Then we need to maximize the margin y = 2/|| w|| or minimize ||w|| subject to constraint (6). This is a quadratic programming (QP) optimization problem that can be expressed as:

$$minimize \frac{1}{2} \|w\|^2 \qquad \text{Eq.(7)}$$

Usually, datasets are often not linearly separable in the input space. To deal with this situation slack variables (ξ_i) are introduced into Eq. (8), where C is the parameter that determines the trade-off between the maximization of the margin and minimization of the classification error. Now the QP optimization problem is given by

$$minimize \frac{1}{2} \|w\|^2 + C \sum_i \xi_i \qquad \text{Eq.(8)}$$

subject to

$$y_i(w.x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \text{ for i=1, 2, } \ldots \qquad \text{Eq.(9)}$$

The solution to the above optimization problem has the form:

$$f(x) = w.\phi(x) + b = \sum_{i=1}^{l} C_i \phi(x_i).\phi(x) + b \qquad \text{Eq.(10)}$$

where

Φ (.) is the mapping function that transforms the vectors in input space to feature space. The dot product in Eq. (10) can be computed without explicitly mapping the points into feature space by using a kernel function. Here the proposed method has used the linear kernel of the form

$$k(x, y) = x^T.y \qquad \text{Eq.(11)}$$

5. EXPERIMENTAL SETUP AND TESTING RESULTS

This section includes the evaluation of the proposed image recaptured detection method. In this experiment, I used smart phone recapture image database proposed by Xinting Gao et al., [8]. The real-scene images are obtained by three popular brands of smart phones including Acer M900, Nokia N95 and HP iPAQ hw6960 which have back-facing camera. The recaptured images are obtained by using three types of DSLR cameras including Nikon D90, Canon EOS 450D and Olympus E-520.

For this experiment, totally 2231 images including 1094 real-scene images and 1137 recaptured images are used. In this experimental setup, three different experimental setups are performed in

order to demonstrate the performance of proposed method. First, the proposed method is compared with several state-of-the-art methods. Second, the performances of different combinations of features used in this paper are compared. Last, the performances of brands of smart phones are compared.

5.1. Experiment I

The performance of proposed method is compared with the state-of-the-art methods. As suggested by Ke et al., [2], the whole dataset is partitioned into training and testing images as shown in Table 3 and measured the performance using accuracy.

Table 3. Training and testing image selection.

Selection	Training set	Testing set
Selection I	70 %	30 %
Selection II	60 %	40 %
Selection III	50 %	50 %
Selection IV	30 %	70 %

Twenty six dimensional low-level features including Texture, HSV colour and Blurriness are extracted from the training images. Table 4 shows the recognition rate of proposed method using different training and testing samples.

Table 4. Recognition rate as accuracy with different image samples on Smart Phone Recapture Image Database [8].

Selection	Accuracy
Selection I	86.67 %
Selection II	86.25 %
Selection III	84.50 %
Selection IV	78.21 %

Table 5 compares the performance of the proposed method with state-of-the-art methods. According to the accuracy that is shown in Table 5, it can be seen that proposed method gives similar performance compared to other methods.

Table 5. Comparison of feature dimension and performance achieved.by different methods on smart phone recapture image database [8].

Features	Dimensions	Accuracy
Physics [1]	166	91.3 %
Wavelet statistics [4]	216	80.67 %
Proposed method	26	86.67 %

5.2. Experiment II

Table 6. Dimensions and performance on the different combinations of features.

Features	Dimensions	Accuracy			
		Section I	Section II	Section III	Section IV
Texture	16	90.00 %	85.00 %	82.00 %	74.29 %
HSV	9	56.67 %	62.50 %	62.00 %	54.29 %
Blurriness	1	63.33 %	55.00 %	49.00 %	50.00 %
Texture + HSV	25	80.00 %	75.00 %	72.00 %	70.00 %
Texture + Blurriness	17	90.00 %	85.00 %	82.00 %	74.29 %
HSV + Blurriness	10	70.00 %	60.00 %	58.00 %	58.57 %

In order to find out the robust feature in recognition of recaptured images, I have computed the performance of recaptured image detection method using several sets of image features such as [texture], [HSV colour], [blurriness], [texture + HSV colour], [texture + blurriness] and [HSV colour + blurriness]. The results are shown in Table 6.

It is observed that the CS-LBP operator which is used to extract the texture feature is the most robust feature in recognition of recaptured images.

5.3. Experiment III

In order to find out the smart phone which has a good quality capturing process, the proposed method is applied with different image samples; those are captured by three popular brands of smart phones including Acer M900, Nokia N95 and HP iPAQ hw6960. Table 7 is concluded that the proposed image recaptured detection method based on physical based features is more effective for Acer M900 than Nokia N95 and HP iPAQ hw6960

Table 7. Performance on the brands of smart phones

Brand	Accuracy			
	Section I	Section II	Section III	Section IV
Acer	81.67 %	81.25 %	74.00 %	72.14 %
HP	68.33 %	66.25 %	63.00 %	70.71 %
Nokia	76.67 %	72.50 %	76.00 %	72.14 %

In this experiment it is concluded that the proposed image recapture detection method has achieved a comparable classification performance on low dimensional features including texture, HSV colour and blurriness. Among them, the texture which is extracted using CS-LBP operator is crucial for the recognition problem and also the proposed method is more effective for Acer M900.

6. CONCLUSIONS

In this paper, I proposed an image recapture detection method based on set of physical based features which uses combination of low-level features including texture, HSV colour and blurriness. This proposed method is efficient with good recognition rate of distinguishing real-

scene images from the recaptured ones. Even though the proposed method possesses low dimensional features, it works excellently in both situations where in less training as well as more training images.

There is a restriction in my research work. This limitation is that the dataset is consisting with images taken only back-facing camera in three types of smart phones as I described in the section Dataset. This will be an effect for the Experiment III to propose the overall performance of brands of smart phones using the proposed method.

The future work is to use the most robust feature to train two dictionaries using the K-SVD approach [13]. Using these two learned dictionaries, we would be able to determine whether a given image has been recaptured. Another work is to extract more other features and measure the performance to find the best combinations of all the features.

REFERENCES

[1] X. Gao, T.-T. Ng, B. Qiu&S.-F. Chang, (2010) "Single-view recaptured image detection based on physics-based features", IEEEInternational Conference on Multimedia and Expo (ICME), pp1469-74.

[2] Y. Ke, Q. Shan, F. Qin & W. Min, (2013) "Image recapture detection using multiple features", International Journal of Multimedia and Ubiquitous Engineering, Vol. 8, No. 5, pp71-82.

[3] H. Yu, T. –T. Ng&Q. Sun, (2008) "Recaptured Photo Detection Using Specularity Distribution", IEEEInternational Conference on Image Processing, pp3140-3143.

[4] H. Farid&S. Lyu (2003) "Higher-order wavelet statistics and their application to digital forensics", IEEE Workshop on Statistical Analysis in Computer Vision.

[5] H. Cao &A. C. Kot, (2010) "Identification of Recaptured Photographs on LCD Screens", IEEEInternational Conference on Acoustics, Speech and Signal Processing, pp1790-1793.

[6] D. Ngo, (2008) Vietnamese security firm:Your face is easy to fake, [Online], Available: http://news.cnet.com/8301-17938105-10110987.html

[7] J. Bai, T.-T. Ng, X. Gao&Y. Q. Shi, (2010) "Is physics-based liveness detection truly possible with a single image?",IEEEInternational Symposium on Circuits and Systems (ISCAS).

[8] X. Gao, B. Qiu, J. Shen, T.-T. Ng, &Y. Q. Shi (2011) Digital Watermarking 9th International Workshop, IWDW Revised Selected Papers, pp90-104.

[9] W. Xiaosheng &S. Junding, (2009) "An effective texture spectrum descriptor",Fifth International Conference on Information Assurance and Security.

[10] M. Heikkila&C. Schmid, (2009) "Description of interest regions with Local binary patterns", Pattern Recognit., Vol. 42, No. 3, pp425-436.

[11] F. Crete, T. Dolmiere, P. Ladret&M. Nicolas, (2007) "The blur effect: perception and estimation with a new no-reference perceptual blur metric", SPIEInternational Society for Optical Engineering.

[12] A. Ramanan, S. Suppharangsan &M. Niranjan, (2007) "Unbalanced Decision Tree for Multi-class Classification", IEEE International Conference on Industrial and Information Systems (ICIIS'07), pp291-294.

[13] T. Thongkamwitoon, H. Muammar & P. L. Dragotti (2014) "Robust Image Recapture Detection using a K-SVD Learning Approach to train dictionaries of Edge Profiles", IEEE International Conference on Image Processing (ICIP), pp5317-5321.

AUTHORS

S. A. A. H. Samaraweera received the B.Sc. Special Degree in Computer Science from University of Jaffna, Sri Lanka in 2016. Her current research interests include image processing and digital image forensic.

B. Mayurathan received PhD Degree in Computer Science from University of Peradeniya, Sri Lanka in 2014. Her current research interests include Computer vision and Machine Learning.

INTENTIONAL BLANK

TOPIC BASED ANALYSIS OF TEXT CORPORA

Madhumita Gupta[1] and Sreya Guha[2]

[1]Palo Alto, USA
madhumita@gmail.com
[2]Palo Alto, USA
sreyaguha@gmail.com

ABSTRACT

We present a framework that combines machine learnt classifiers and taxonomies of topics to enable a more conceptual analysis of a corpus than can be accomplished using Vector Space Models and Latent Dirichlet Allocation based topic models which represent documents purely in terms of words. Given a corpus and a taxonomy of topics, we learn a classifier per topic and annotate each document with the topics covered by it. The distribution of topics in the corpus can then be visualized as a function of the attributes of the documents. We apply this framework to the US State of the Union and presidential election speeches to observe how topics such as jobs and employment have evolved from being relatively unimportant to being the most discussed topic. We show that our framework is better than Vector Space Models and an Latent Dirichlet Allocation based topic model for performing certain kinds of analysis.

KEYWORDS

Text Analysis, Machine Learning, Classification

1. INTRODUCTION

In recent years, researchers in fields as diverse as biology, law and the social sciences have started using computational models to analyze corpora of scientific papers, judgments, speeches, etc. These models have enabled researchers to discern patterns and gain insights into their respective fields. For example, such models have been used to corroborate the authorship [6] of several of the Federalist Papers, a collection of 85 essays that promoted the ratification of the constitution of the United States, written by Alexander Hamilton, James Madison, and John Jay, under the pseudonym Publius. The computational analysis of these essays has led researchers to dispute the authorship of 11 of them.

Computational models of text corpora aim to find patterns that capture underlying phenomena in the domain discussed by the documents. A prominent feature of the approaches that are used today is that they are bottom up. As pioneered by Salton et. al. [15] in the Information Retrieval community, the representations, learnt concepts, etc. are all built up purely from the words in the document. As these approaches are driven only by the words in the document, they can be applied to a new corpus without any manual pre-processing { a huge advantage. This generality,

Jan Zizka et al. (Eds) : ACSTY, NATP - 2016
pp. 33– 47, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61403

however, can also be a shortcoming. Many interesting questions cannot be expressed purely in terms of the words that appear in the documents. Consider, for example, the corpus of US Presidential State of the Union speeches (a State of the Union speech is an annual speech given by the President of the United States). In the early years of the US, Native American relations were talked about more frequently than they are today. If we wanted to understand how this is reflected in these speeches, or when this change started, we would need a model that provides "Native Americans" as a model feature, which a pure word based model would not.

Every field has concepts/topics that are central to the discourse in that field and many important questions are in the vocabulary of these concepts/topics (henceforth, referred to as topic). The above question, for example, needs to reference the concept/topic "Native American". While the words in the document do capture the topic(s) discussed, it is difficult to express certain questions only in terms of the words in the document. This is especially the case when:

- Different combinations of words can be used to refer to the same topic. E.g., Sometimes, Native American relations are discussed within the context of certain tribes, requiring the model to be able to recognize the topic from the tribe names.

- Terminology evolves over time. Over the years, Native Americans have been referred to by other names, such as "Indians", a term which is now used to refer to a different community.

- Different participants use different words to talk about the same topic. For example, for the topic "Abortion", different parties use different terms. Democratic Party speakers tend to use the phrases such as "access to contraceptives," and, "womans right to choose," whereas Republican Party speakers tend to use "rights of the unborn," and, "right to life," and so on.

In this paper, we propose a framework for addressing the mismatch between the queries researchers want to ask and the vocabulary of the modeling tools. We are given a corpus of documents, each with a set of attributes, and a taxonomy of topics. Assuming that a single paragraph in a document is restricted to one topic, we build a set of classifiers, one for each topic such that each paragraph is labeled as belonging to one or zero topics. We use our framework to analyze the distribution of topics discussed as a function of the attributes of the documents. We apply our framework to two corpora:

1. The State of the Union Speeches, from 1790 to 2016 (226 speeches),
2. Speeches from the primary and general elections in the presidential elections from the years 1996 to 2016 (17,718 speeches)

Each speech has the date of the speech, a title and the speaker. These two corpora are part of a larger corpus of 19,572 political speeches obtained from [20], [16] and [4]. We created a simple taxonomy of 26 topics corresponding to the most popular issues in political discourse, such as the economy, human rights, etc. (see figure 1 for the full taxonomy). Given the nature of political speeches and the significant time frame over which these speeches were made, there is a wide variation in the words used to discuss a given issue, both across speakers and over time. Though the list of topics in our taxonomy is by no means exhaustive, we believe that it is adequate in size and variety to demonstrate the benefits of our approach.

2. METHODOLOGY

Our goal is to develop a framework with which we can better understand the distribution of a set topics across the documents in a given corpus, as a function of the attributes of the documents. We develop the framework and apply it to the speeches in two corpora mentioned earlier. We analyze the distribution of these topics across these speeches as a function of the speaker, speaker affiliation, year, etc. We do this using not only our proposed framework but also using two popular techniques for quantitative analysis of text corpora. We now discuss these two techniques before describing the details of our framework.

2.1. Existing Models

The models described below are currently the two most widely used models:

Bag of Words Model: Each document is modeled as bag of words, where a "Term Frequency Inverse Document Frequency" (TFIDF) score is computed for each word document pair. This measures the number of times the word occurs in the document, normalized by the frequency of occurrence of the word in the corpus as a whole. The most significant terms in a document can be said to capture the main points in the document. The most widely used application of this model is document search (e.g. Web search).

Latent Semantic Models: Starting with the work by [2], a variety of "latent semantic" models have tried to create more abstract, implicit representations of meaning, which capture the fact that different combinations of words can be used to express the same concept. Latent Semantic Indexing has been used in a variety of applications such as patent discovery, document classification and determining authorship of documents. However, the resulting implicit representations are hard to interpret and words with multiple, evolving meanings cannot be disambiguated easily. Recently, there has been work in machine learning, under the term "Topic Modeling" that computes a generative model for a corpus of documents. Documents are assumed to be generated, according to some distribution (typically, a Dirichlet distribution) from a set of latent topics. The goal is to generate these topics, which should provide insight into what the corpus is about. Though latent semantic approaches such as topic models do try to go beyond words, since their representation of semantics is "latent", it is hard, if not impossible to express questions about a particular topic that is of interest to us.

2.2. Classification

Each speech was divided into several "documents", i.e. paragraphs. Short paragraphs with fewer than 50 words were merged to create bigger paragraphs. A single paragraph is henceforth the equivalent of a document for our classification purposes. Each paragraph may discuss one of our given topics. Note that there will be many paragraphs that don't discuss any of our given topics. One of our key technical problems is to identify when a paragraph is discussing one of the given topics. We do this by using machine learning to build automated classifiers for each topic.

In this work, we build classifiers only for the leaf nodes in our taxonomy. The interior nodes are assumed to be unions of their child nodes. The framework we develop can easily be extended to taxonomies that don't make this assumption. We build our classifiers against the combined

corpus, i.e., there is a single classifier for a topic such as 'Immigration', not one for each of our corpora.

2.2.1. Limitations of Pure Keyword Matching

One very simple way of matching paragraphs with topics is by keyword matching. For example, a topic such as "Jobs/Employment" could be associated with a keyword 'jobs' and every paragraph that contains the word 'jobs' (or a stemmed version of 'jobs') can be associated with this topic. Unfortunately, this method has significant limitations. The same phrase can be used in different contexts with different meanings. For example, the phrase "right to choose" is often used to discuss "Abortion". However this phrase is also used in many political speeches on democracy, to discuss the right of the Vietnamese or the Iranians to choose their own government. Similarly, using jobs as an indicator of a document that matches the topic of Jobs/Employment leads to large numbers of spurious matches (let them do their job, the book of Job, and so on).

Table 1: Comparison of terms used to discuss Nuclear Weapons in 1950's vs in 2010's

Nuclear Weapons in 1950's	atom, missile, communist, soviet, bomb
Nuclear Weapons in 2010's	korea, israel, sanction, iran, deal

Vocabulary changes over time, and the manner in which words are used changes over time. For example, the term "Drone" in today's context typically refers to remote piloted aerial vehicles. However, it also appears in speeches by George Washington to refer to something quite different. A striking example of the changing vocabulary is shown in table 1, which illustrates the change in some of the terms associated with Nuclear Weapons.

Support Vector Machines [9] and other machine learned classifiers, on the other hand, don't rely on simple keyword matches. By using more complex functions that combine partial support for a given topic from different words in the document, they get around some of the limitations of simple keyword matches.

2.2.2. Creating Training Data

In order to learn a classifier, we need a set of labeled examples. We use the following procedure to create a training set.

1. For each topic, we manually specify a set of phrases that, with high probability, identify paragraphs about that topic (Table 2).

2. For each phrase, we extract the paragraphs in the corpus containing that phrase, giving us a set of paragraphs for each topic. For some topics, for example, Drugs, we had as few as 300 matches in all. For others, such as Jobs/Employment, we had 13,000 matches.

3. We manually check a small sample of the paragraphs to check whether the paragraph corresponds to the topic. The correct ones go into the positive training set and the (small number of) wrong ones go into the negative training set.

4. We extract top K ($\approx$ 10) words/phrases in these positive training set that have a high TFIDF. These are added to the list of phrases for the issue.

5. We extract additional matches using these new phrases and repeat the labelling into positive and negative training sets.

6. The negative training set for each issue is augmented with samples from the positive training set of other issues.

For each training set, we took positive and negative examples in the ratio of 1:4, i.e., about 250-500 positive examples, and 1000-2000 negative examples.

2.2.3. Pre-processing Data

Each paragraph is pre-processed using the statistical package R [14] as follows.

1. The text is lower-cased, punctuation and whitespace removed.

2. All numbers and stop words are removed (words such as a, is, after, before, etc. that are very frequent, but do not provide any information).

3. The words are stemmed (reduced to their base form).

4. Finally, each paragraph is broken into a set of "features" - the single word phrases (also known as unigrams) it is composed of.

These paragraphs and features are then composed into a "Document Term Matrix" (dtm): a matrix whose rows are the paragraphs and columns are the union of all features across all paragraphs. The values are the TFIDF scores of the features:

Table 2: Sample of phrases used to create training set.

Topic	**Keywords / Phrases**
Jobs/Employment	middle class, GDP, unemployment, recession, job creation, jobless, great depression, economic recovery, minimum wage
Immigration	Anchor babies, illegal immigration, H1B, deportation, border guard
Trade	tariff, laissez faire, nafta, free trade, tpp, import duty
Taxes	income tax, death tax, redistribution, 1 percent, trickle down, tax cut
Healthcare	medicare, medicaid, obamacare, medical insurance, public option
Social Security	retirement age, safety net, social security fraud, government handout
Nuclear Weapons	thermonuclear, ballistic missiles, arms limitation, mutually assured destruction, north korea, kim jong un, nuclear disarmament, plutonium
Climate Change	rising temperatures, global warming, fracking, carbon, Kyoto, Paris Climate Agreement, Fossil Fuel, sea level rise, glacier, greenhouse
Race Relations	separate but equal, racism, Brown v. Board of Education, Martin Luther King, voting rights act, desegregation, negro, emancipation, slavery
Drugs	just say no, war on drugs, heroin, cocaine, drug rehabilitation, opioid, overdose
Terror	Al Qaeda, Taliban, bin laden, ISIS, drones, 911, twin towers, world trade center, benghazi, Tora Bora, Hizballah
Inflation	bimetallism, gold standard, de ation
Native American	indian, apache, cherokee, western settlements, indian affairs, cheyennes

$$\text{dtm}[i,j] = (\text{the number of times feature } j \text{ occurs in paragraph } i) \times \log_2\left(\frac{\text{total number of paragraphs}}{1 + \text{total number paragraphs in which feature } j \text{ occurs}}\right)$$

2.2.4. Tuning the pre-processing

There are many parameters that can be tuned during the pre-processing, each of which can affect the classification. Some of these include:

- **Stop Words:** TFIDF based scoring alone is not enough to eliminate the effect of very frequent terms. The programming package we used (tm for NLP) removes a standard set of stop words such as 'the', 'and', etc. However, there are many other words such as "will", "campaign", "political", which occur very frequently in this corpus, which do not provide any significant information. The TF-IDF measure for scoring features is intended to reduce the importance of exactly such words. However, they are so frequent within the documents, that merely counting the documents that they occur in does not have a significant enough impact on their score, and they are not dropped from the feature matrix, and sometimes end up as or more important than high information words. As an example, in a set of 7980 documents used to train the concept of Abortion, "will" occurred 4551 times in 2454 documents. "Partial birth" on the other hand occurred 18 times in 17 documents. The resulting IDF makes "partial birth" 9 times more significant as compared to "will", but the high TF of "will" counteracts IDF, and "partial birth" and "will" end up with similar TFIDF scores.

- **Stemming Variants** of a word (legislation, legislate, etc.) should usually be treated similarly for purposes of classification. We therefore 'stem' each word occurrence to a root and treat them all similarly. However, occasionally, stemming can change its meaning. For example, for the topic "Jobs/Employment", for the training set, we looked for documents containing "recession". The stemmed form of "recession" is "recess", and we had many spurious matches such as "Congress is in recess". However when we trained without stemming, the results were slightly worse - e.g. for Safety the F-measure went from 92.85% to 91.11%, for Social and Health from 76.45% to 74.33% while for Human Rights it went from 85.78% to 87.41%. So, despite losses such as "recession", we continued using stemming.

- **Unigrams, bigrams, etc.** In addition to unigrams, we experimented with bigrams and trigrams (two word and three word phrases). Example bigrams and trigrams are "Star Wars" and "National Reconstruction Act". If we use only bigrams and trigrams, our specificity increases to a point where we are left with too few matches, and if we use both unigrams and bigrams, the number of unigrams swamp out the bigrams. So, we used only unigrams.

- **Imbalanced Data** Our training data set had fewer positive examples for each topic as compared to negative examples. As a result, the initial classifiers labeled all documents as negative (not matching the topic). We solved this problem by assigning weights to the positive and negative examples, in inverse proportion to the number of examples of that type.

2.2.5. Learning Algorithms

With the Document Term Matrix as input, we used three different algorithms to learn classifiers for each topic: logistic regression [10], Classification and Regression Trees (CART)[1], and Support Vector Machines (SVM) [9]. For each algorithm, we measured its precision on a hold back from our training set. Of the three, Support Vector Machines with a radial basis kernel [18] were the most accurate.

3. EXPERIMENTAL RESULTS

We had a total of 19572 political speeches, made by 245 speakers. Our taxonomy (Fig. 1) has 26 topics with an average of 11 phrases per topic for creating the initial training set. We used the tm (Natural Language Processing, [5]), XML ([11]) and plyr ([19]) packages of the R [14] programming language to preprocess the data and then to convert it into the document term matrix. For classification, we used the randomForest ([12]) and rpart ([17]) packages for building the CART classifiers, and the e1071 package ([3]) for building Support Vector Machine (SVM) classifiers with linear and radial basis kernels. We obtained the best results with SVM with a Radial Basis kernel.

As with all classifiers, we have to trade off precision (the fraction of sentences classified as belonging to an issue that are indeed of that issue) with recall (the fraction of all sentences of an issue that are identified as being of that issue). We used the F-measure, the harmonic mean of precision and recall to estimate the performance. The precision and recall numbers used in the tuned classifiers, for a sample of topics, are given in table.

Topic	Precision	Recall	F-measure
Safety	93.42%	92.28%	92.85%
Social & Health	81.87%	71.71%	76.45%
Human Rights	83.07%	88.68 %	85.78%
Economy	64.0%	66.36%	65.16%
Defense & Foreign Relations	74.72%	72.15%	73.41%

To compare our framework with word based techniques, we also analysed our corpora both by modeling each document as a Bag of Words and by using Topic Modeling to identify a set of topics (that the corpora are 'about'.) The words with the highest TFIDF from the speeches of a selected set of speakers in the two corpora are given in tables 3 and 4.

We used two different Topic Modeling tools ([13] and [8]) to generate topics for our combined corpus (remember that build our classifiers against the combined corpus, so that there is a single classifier for a topic such as 'Immigration' that works for all speeches). The results generated by [13] were better and are given in table 5. As we can see, the automatically generated topics are not easily understandable. In particular if we are trying to answer questions such as "how did discussion about `Native Americans' in the State of the Union speeches evolve", since none of the topics corresponds to `Native Americans', it is difficult to frame the question in terms of the automatically generated topics of table 5.

Table 3: Words with highest TFIDF used by Presidents in State of Union Speeches

George Washington	indian, treaty, militia, session, tribes, rendered, hostile, cherokee, frontier, tranquility, deliberations, orins, insurrection, expedient, attaching, observed
Andrew Jackson	treaty, treasury, indian, session, deposits, france, exercise, minister, vessels, branches, deemed, payment, portion, intercourse, rendered, ports, possessions
Abraham Lincoln	slavery, territorial, nebraska, missouri, compromise, emancipation, negroes, clay, repealing, prohibition, framed, douglas, ordinance, rebellion, sacr
Theodore Roosevelt	panama, interstate, island, navy, forests, canal, wageworker, republic, isthmus, railroad, exercise, treaty, philippine, territorial, supervised, tariff
Franklin D. Roosevelt	nazis, germany, axis, planes, pacific, farmers, japan, sea, material, british, hitlerism, italy, britain, island, agriculture, recovery, continent
Ronald Reagan	soviet, in ation, missile, lebanon, strategic, nicaragua, laughter, gorbachev, grenada, revolution, space, treaty, sdi, israel, rgeneva, sandinistas, totalitarian
Bill Clinton	medicare, police, bridges, bosnia, guns, somalia, russia, bipartisan, somalis, brady, laughter, kosovo, covenants, nato, cold, doctorate, black, teaches

Table 4: Words with highest TFIDF used in Presidential primary speeches

Ted Cruz	pastor, churches, houston, baptist, trump, cochair, rubio, activist, religious, christian, rep, coalition, polk, tea, texas, islamic, marriage, ministries
Bernie Sanders	billionaires, vermont, climate, burlington, saturday, superpac, sunday, isis, inequalities, donald, friday, turnout, weaver, mondays, minimum
Rick Santorum	verona, obamacare, romneycare, marriage, rpa, gingrich, abortions, hogan, mandates, healthcare, bailouts, newt, radical, contrast, repealing, tea
Hillary Clinton	Activist, mortgages, nevada, foreclosure, coverage, manchester, rural, nurse, guns, des moines, lgbt, latino, healthcare, hispanic, longterm, treatment, green
Donald Trump	illegal, grafton, georgia, merrimack, hillsborough, cheshire, israel, falwell, tremendous, lewandowski, cochair, belknap, palin, patrol, ballot
John McCain	Palin, acorn, usn, biden, admiration, anncr, arlington, ayers, coal, surge, drilling, afghanistan, blogs, usaf, withdrawal, iraqi, abc, mortgages
Mitt Romney	medicare, michigan, Ryan, obamacare, illegal, huckabees, recovery, journal, bain, editorial, marriage, fox, giuliani, newt, biden, veto, ret, solyndra

4. DISCUSSION

One of the primary motivations in creating this framework is to enable us to analyze how the discourse (in a given corpus) has evolved with respect to a given set of topics. This kind of analysis often facilitates better understanding of the underlying phenomena.

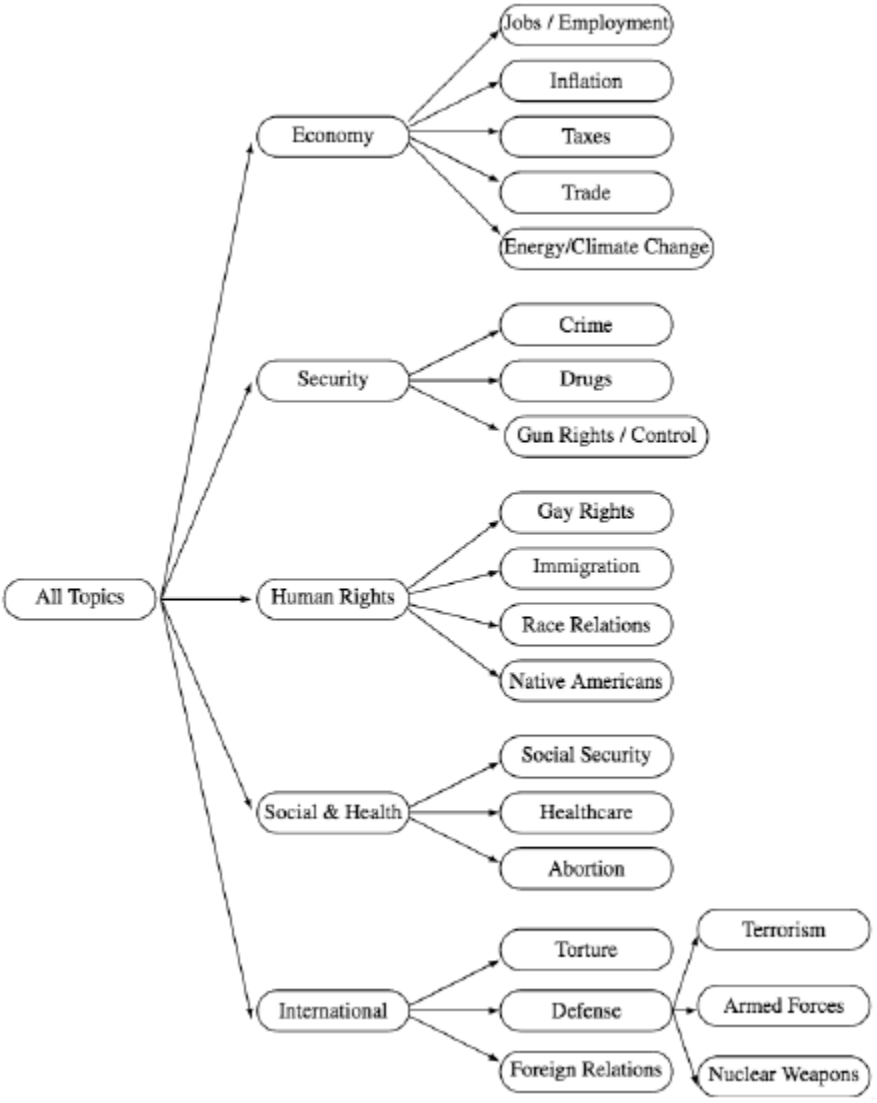

Figure 1: Topic Hierarchy

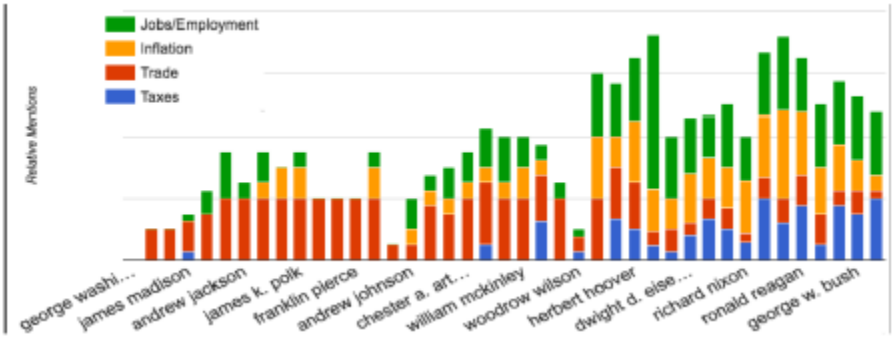

Fig 2 : Relative mentions of the 'Economy' topic in State of Union Speeches

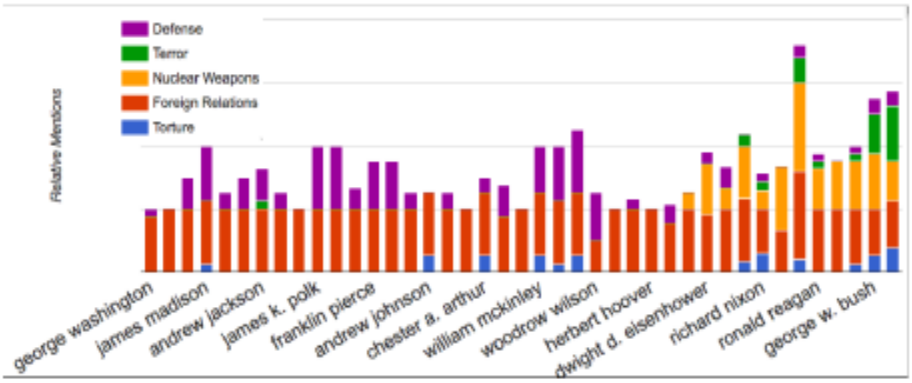

Fig 3: Relative mentions of the 'International' topic in State of Union Speeches

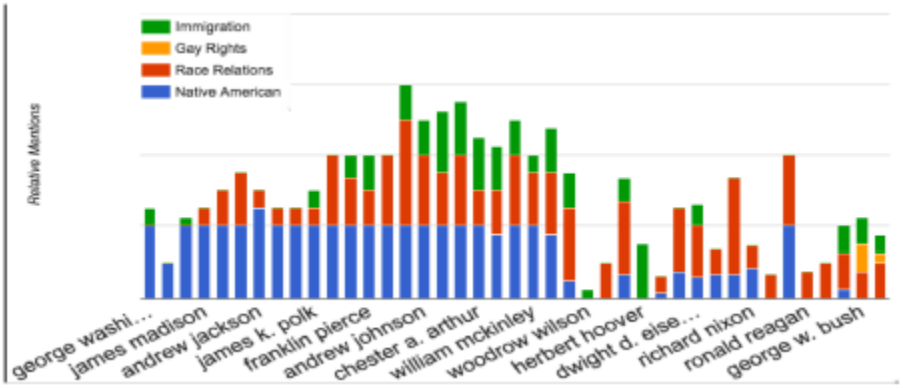

Fig 4: Relative mentions of the 'Human Rights' topic in State of Union Speeches

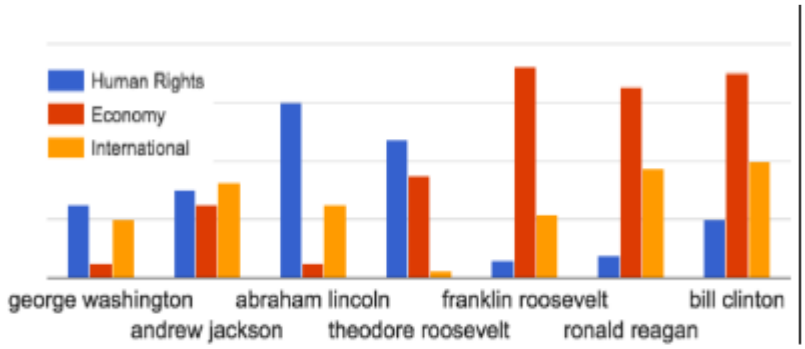

Fig 5: Change in relative coverage of topics over time in State of Union Speeches

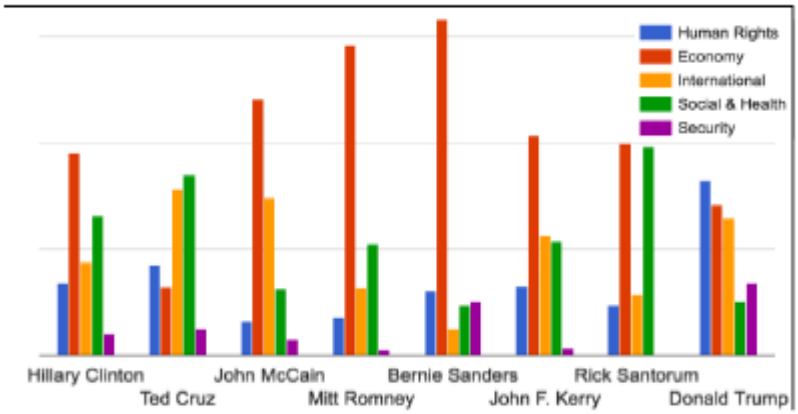

Fig 6: Relative coverage of topics in primary election speeches

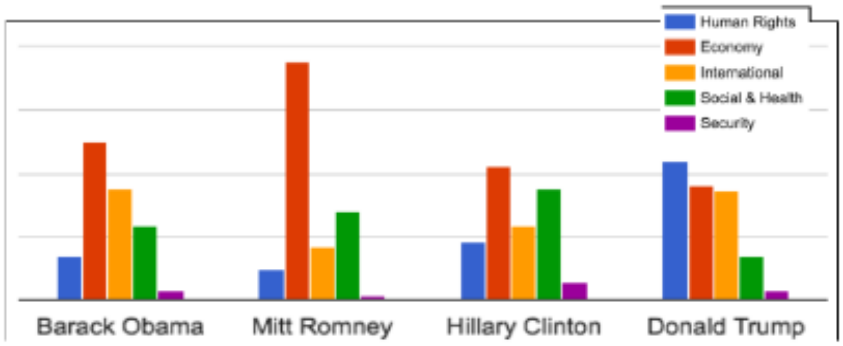

Fig 7: Relative coverage of topics in general election speeches

Table 5: Autogenerated topics using LDA

Topic 1	government, american, country, america, united, promise, times, states, new
Topic 2	santorum, state, rick, government, vote, united, romney, president, life
Topic 3	tax, plans, jobs, romney, percent, federal, energy, taxes, American
Topic 4	government, states, united, congress, year, 000, law, people, public
Topic 5	santorum, great, wall, shall, country, present, people, government, time
Topic 6	santorum, congress, public, life, campaign, war, rick, romney, said
Topic 7	public, government, congress, country, law, 000, people, shall, American
Topic 8	perry, romney, rick, gov, president, state, santorum, said, America
Topic 9	world, people, america, life, government, great, country, peace, time
Topic 10	search, romney, public, health, year, massachusetts, care, choose, month

The output of the analysis is a database of paragraphs with annotations for topic, speaker, date, context (i.e., primary election, general election or State of Union). We use the graphing functions provided by Google Sheets [7] to help us detect patterns in the data. In this section, we present some insights/observations as illustrated by the relevant graphs. We also use the bag of words model to generate a list of the words with the highest TFIDF scores and compare the two approaches from the perspective of their ability to enable us to make these observations.

4.3. Evolution of Topics

Figures 2, 3 and 4 plot the number of mentions of the topics `Economy', `International' and `Human Rights', in the state of union address by each president. Since the number of state of union speeches given by a particular president varies from 1 to 13, we normalize the number of mentions by the number of speeches given by that president. Looking at these graphs, we make the following observations

- Economy is the biggest issue nowadays, but this was not always the case. Economy as the theme of the presidency seems to started at around the time of the Great Depression. Further looking at figure 2 we see the mixture of topics related to the Economy has changed. In the earlier days of the Union, the Economy related discussions were more around the topic of `Trade'. The topic of 'Jobs / Employment' has become an increasingly important part of the discussion in the last 80 years. Taxes start coming up repeatedly only around 1900, when income tax was first introduced.

- In figure 3, we see that Foreign Relations and Defense are a constant theme, but we see the recent emergence of subtopics such as Terror.

- Looking at figure 4, we see a recent fall in discussions about Native Americans. The issue of Gay Rights is relatively recent and has made its appearance in State of Union speeches only in the last few presidencies. It is interesting to see that race relations have always been a topic in these speeches. Indeed, comparing figures 3 and 4, we see that Foreign Relations and Race Relations are the only two topics that have figured prominently in over 90% of the speeches.

- We see that different topics have very different temporal behaviors. Some (such as "Race Relations") have maintained a relatively steady presence through the history of these speeches. Others, such as "Native American" have remained steady for a period of time and then declined. Some topics such as the "Jobs/Employment", which are now central, took a long time to develop. Some other issues, such as "Nuclear Weapons" sprung into prominence relatively fast.

4.4. Relative Coverage of Topics

The relative coverage of topics gives us insight into the changing priorities of a community.

Graph 5 shows the relative coverage of topics across seven presidents drawn from different time periods. As we can see, the biggest issues in the early years of the country were "Human Rights" (actually, "Native American" issues) and "Foreign Relations", and now it is "Economy". Discussions about the topic "Foreign relations" have stayed relatively the same throughout time; however, coverage of the topic "Jobs/Employment", and of the topic "Economy" overall have significantly increased since the Great Depression. Comparing graph 5 to the TFIDF table 3 for the same presidents, we see that the analysis produced by our framework is substantially better for answering these kinds of questions.

Graph 6 shows the relative coverage of topics for a number of recent presidential candidates. We see that "Economy" is the most important topic for most, but not all of them. The graph clearly

shows which issues were most important to each candidate. Graph 7 shows the relative coverage of topics during the general election. Comparing graphs 6 and 7, we can see that the coverage of topics during general elections is more even than during the primaries.

4.5. Word Based Analysis

As can be seen from table 3, the words used during different time periods have evolved significantly. It is interesting to see how well the words with the highest TFIDF capture the major issues of each presidency. For example, the most significant words for Lincoln include 'slavery', 'rebellion' and 'emancipation' and the most significant words for Franklin Roosevelt include 'nazis', 'japan' and 'recovery'. The significant words from the primary speeches on the other hand do not exhibit such clarity. Despite the words in table 3 accurately reflecting the major topics of each presidency, we can see that would be difficult to draw conclusions such as those enabled by figures 2, 3, 4 from these word lists.

5. CONCLUSIONS AND FUTURE WORK

We introduced a framework for analysing the documents in a corpus, from the perspective of a given taxonomy of topics. The framework facilitates annotation each document with the topics covered and subsequent analysis of the distribution of topics as a function of the attributes of the documents. We implemented this framework and applied it to to two corpora of political speeches. We showed how the framework enables us to draw insights into into the relative composition of speeches and to the evolution of topics.

There are many different areas of future work. Looking at the topic hierarchy (fig. 1), we can see that it is hard to organize topics into a clean tree. Some topics, like "Climate Change", legitimately belong in multiple higher level categories | "Economy", "International", etc. We would like to extend our approach to taxonomies that are allow for multiple parents for each topic. We have assumed that the topic hierarchy is relatively static. However, just as the terminology used to discuss a topic evolves, the topic hierarchy itself evolves. E.g., In the early years, "Native American" issues were probably more part of "Defense" than "Human Rights". Over time, they have become more part of "Human Rights". Handling evolving taxonomies is another direction for future work.

The biggest limitation of our approach is the cost of building classifiers. A taxonomy with thousands of topics could be very expensive to build classifiers for. As we can see from table 3, the high TFIDF words do sometimes capture the main points of a document/ speaker. Though words by themselves are inadequate for capturing more abstract concepts like `Race Relations' they are good at capturing more specific topics like `emanicipation proclamation'. One line of future work involves combining the two approaches, wherein for the more general topics, we use machine learnt classifiers, but as the topics get more specific, a combination of topic modeling and word vector based approaches can be used to fill out the taxonomy.

REFERENCES

[1] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. CRC press, 1984.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391, 1990.

[3] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, and M. F. Leisch. Package e1071. R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html, 2009.

[4] M. E. Eidenmuller. American Rhetoric: The Power of Oratory in the United States. http://www.americanrhetoric.com, 2001-2016. [Online; accessed 7-July-2016].

[5] I. Feinerer and K. Hornik. Text mining package, 2015.

[6] G. Fung. The disputed federalist papers: Svm feature selection via concave minimization. In Proceedings of the 2003 Conference on Diversity in Computing, pages 42{46. ACM, 2003.

[7] Google.com. Google Sheets : Create and edit spreadsheets online. https://www.google.com/sheets, 2016.

[8] B. Grun and K. Hornik. topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13):1{30, 2011.

[9] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. IEEE Intelligent Systems and their Applications, 13(4):18{28, 1998.

[10] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. Applied logistic regression, volume 398. John Wiley & Sons, 2013.

[11] D. T. Lang. Xml: Tools for parsing and generating xml within r and s-plus. R package version, pages 3{9, 2012.

[12] A. Liaw and M. Wiener. Classification and regression by randomforest. R News, 2(3):18{22, 2002.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R.Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825{2830, 2011.

[14] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[15] G. Salton and M. J. McGill. Introduction to modern information retrieval. McGraw-Hill, Inc., 1986.

[16] The Miller Center. American President: A Reference Resource. http://millercenter.org/president, 2016. [Online; accessed 7-July-2016].

[17] T. M. Therneau, B. Atkinson, and B. Ripley. rpart: Recursive partitioning. r packageversion 3.1{42. Computer software program retrieved from http://CRAN. R-project.org/package= rpart, 2010.

[18] J.-P. Vert, K. Tsuda, and B. Sch olkopf. A primer on kernel methods. Kernel Methods in Computational Biology, pages 35{70, 2004.

[19] H. Wickham. The split-apply-combine strategy for data analysis. Journal of Statistical Software, 40(1):1{29, 2011.

[20] J. Woolley and G. Peters. The American Presidency Project. http://www.presidency.ucsb.edu/index.php, 2009-2016. [Online; accessed 7-July-2016].

INTENTIONAL BLANK

DICTIONARY BASED AMHARIC-ARABIC CROSS LANGUAGE INFORMATION RETRIEVAL

H L Shashirekha[1] and Ibrahim Gashaw[2]

Department of Computer Science, Mangalore University,
Mangalagangotri, Mangalore-574199
[1]hlsrekha@gmail.com
[2]ibrahimug1@gmail.com

ABSTRACT

The demand for multilingual information is becoming perceptive as the users of the internet throughout the world are escalating and it creates a problem of retrieving documents in one language by specifying query in another language. This increasing demand can be addressed by designing automatic tools, which accepts the query in one language and retrieves the relevant documents in other languages. We have developed prototype Amharic-Arabic Cross Language Information Retrieval System by applying dictionary-based approach that enables the users to retrieve relevant documents from Amharic-Arabic corpus by entering the query in Amharic and retrieving the relevant documents both Amharic and Arabic.

KEYWORDS

Information Retrieval, Dictionary, Machine Translation, Relevance Feedback.

1. INTRODUCTION

With the rapid growth of the Internet, the World Wide Web (WWW) has become one of the most popular medium for spreading multilingual information. The need for multilingual information is becoming perceptive as the users of the internet throughout the world are ever increasing. This ability to disseminate multilingual information has increased the need to automatically intervene across multiple languages, and in the case of the WWW, access to "foreign language" Web pages [1]. The increasing necessity for retrieval of multilingual documents opens up a new branch of Information Retrieval (IR) called Cross Lingual Information Retrieval (CLIR) [2]. Its goal is to accept information, transform it into a searchable format and provide an interface to allow a user to search and retrieve information in different languages [3]. CLIR has lot of applications, such as adhoc retrieval, text summarization, question answering, and text classification to ensure maximal accessibility to digital repository for much wider audience [4].

In addition to the challenges of conventional IR, CLIR systems possess lot of challenges related to language issues [5], such as;

Jan Zizka et al. (Eds) : ACSTY, NATP - 2016
pp. 49– 60, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61404

a. Translation disambiguation, due to homonymy and polysemy [6] creates problems to find the most appropriate translation for a given word

b. Lacking appropriate resources for evaluations of CLIR with low density languages

c. Inflection words in the query cannot be easily located as translated root words in the dictionary, due to stemming

d. New words get added to the language which may not be recognized by the existing system, resulting in out of vocabulary (OOV) and

e. Most of OOV words such as technical terms and named entities in the query reduces the performance of the system

According to Cardenosa et.al, [5], CLIR approaches can be categorized into three; Document translation, Query translation, and Interlingua translation.

- **In document translation,** every document has to be translated into the query language and then retrieval will be performed using classical IR techniques. It can be applied offline to produce translations of all documents well in advance and offers the possibility to access the content in his/her own language. However, machine or (large scale) human translation may not always be a realistic option for every language pair as it is time consuming since every document needs to translated to other languages irrespective of their usage.

- **Query translation approach** is the translation of query terms from source language to the target language. In this approach online translation can be applied to the query entered by a user and it is possible for a user to reformulate, elaborate or narrow down the translated query. Translating a query by dictionary look-up is far more efficient than translating entire document collection. However, it is unreliable since short queries do not provide enough contexts for disambiguation in choosing proper translation of query words and does not exploit domain-specific semantic constraints and corpus statistics in solving translation ambiguity.

- **In Interlingua translation** approach, the source language, i.e. the text to be translated is transformed into an Interlingua, i.e., an abstract language-independent representation. The target language is then generated from the Interlingua. This approach is useful if there are no resources for a direct translation but it has lower performance than direct translation.

Translation techniques in CLIR are categorized into direct and indirect translation [7]. Direct translation uses Machine Readable Dictionary (MRD), parallel corpora, and machine translation algorithm or in combination.

- **In Dictionary based translation** the query words are translated to the target language using MRD [8]. MRDs are electronic versions of printed dictionaries, and may be general dictionaries, specific domain dictionaries, or a combination of both. It has been adopted in CLIR because bilingual dictionaries are widely available.

- **Parallel corpora** contain a set of documents and their translations in one or more other languages. These paired documents can be used to meet the most likely translations of terms between languages in the corpus.

- Query translation can be implemented by using a **Machine Translation** (MT) system to translate documents in one languages in the corpora into the language of a user's query which can be done offline in advance or online [9].

Indirect translation is a common solution when there is an absence of resources supporting direct translation. It can be applied by transitive or dual translation system. In case of transitive translation, the use of an intermediary (pivot) language, which is placed between the source query and the target document collection, is used to enable comparison with the target document collection. In the case of dual translation systems, both the query and the document representations are translated into the intermediate language [10].

In all the above-mentioned cases, a key element is the mechanism to map between languages. This translation knowledge can be encoded in different forms as a data structure of query and document-language term correspondences in a MRD or as an algorithm, such as a machine translation or machine transliteration system [11]. While all of these forms are effective, the latter require substantial investment of time and resources for the development and it is not widely or readily available for many language pairs.

CLIR is becoming a promising field of research which bridges the gap between different languages and hence between different people speaking different languages and of different culture. As CLIR is in its infancy, many works related to many language pairs are attempted. Amharic-Arabic is one such language pairs which needs to explore for CLIR.

According to the 2007 census, Amharic speakers encompass 26.9% of Ethiopia's population. Amharic is also spoken by many people in Israel, Egypt and Sweden [1]. Arabic is a natural language spoken by 250 million people in 21 countries as the first language, and Islamic countries as a second language [8]. Ethiopia is one of the countries, which have more than 33.3% of the population who follow Islam, and they use Arabic language to teach religion and for communication purpose. The Arabic and Amharic languages belonging to the Semitic family of languages [12], where the words in such languages are formed by modifying the root itself internally and not simply by the concatenation of affixes to word roots. Amharic and Arabic are very rich morphology languages.

The current Amharic writing system consists of a core of thirty-three characters (ፊደል, fidel) each of which occurs in a basic form and in six other forms known as orders [1]. The non-basic forms are derived from the basic forms by more-or-less regular modifications. Thus, there are 231 different characters. The seven orders represent syllable combinations consisting of consonant and following vowel. This characteristic according to Abebayehu [13], makes the Amharic writing system a syllabic writing system. A character or a symbol is used to represent a phoneme, which is a combination of a vowel and a consonant. These are written in a unique script that is now supported in Unicode (U+1200 - U+137F) [14].

The Arabic alphabet consists of 28 characters or 29 characters if the Hamza is considered as a separate character. It is written from right to left like Persian, Hebrew, unlike many international languages. Three of the Arabic characters appear in different shapes as follows [15][16]:

- Hamza (ء) is sometimes written :ا, إ or أ (alif)
- Ta marbouta (ة) like t in English found atthe end without two dots (o = ha)
- Alifmaqsurah (ى) is the character (ي=ya) without dots.t

The above three characters pose some difficulties in the setting up a CLIR system. Some of Arabic language resources ignore the Hamza and the dots (.) above "ta marbouta" to unite the input and output for these characters. In Arabic there is a whole series of non-alphabetic signs, added above or below the consonant letters to make the reading of the word less ambiguous.

Both Arabic and Amharic languages possess translation challenges for many reasons [17][18]; such as Arabic sentences are usually long and punctuation has no or little effect on interpretation of the text. Contextual analysis is important in Arabic and Amharic in order to understand the exact meaning of some words. For example, in Amharic, the word "ገና" can have the meaning of Christmas holiday or waiting something until it happens. Characters are sometimes stretched for justified text, which hinders the exact much for same word. In Arabic, synonyms are very common. For example, "year" has three synonyms in Arabic عَام، حول، سنة and all are widely used in every day communication. Another challenge in Arabic is the absence of discretization (sometimes called vocalization). Discretization can be defined as a symbol over and underscored letters, which are used to indicate the proper pronunciations as well as for disambiguation purposes. The absence of discretization in Arabic texts poses a real challenge for Arabic natural language processing, As well as for translation, leading to high ambiguity. Though the use of discretization is extremely important for readability and understanding, they don't appear in most printed media in Arabic regions nor on Arabic Internet web sites. They are visible in religious texts such as Quran, which is fully discretised in order to prevent misinterpretation.

Ethiopia has good socio-economic relationships with Arabic countries; they are communicating using the Arabic and Amharic languages. For example, reports sent between Ethiopia and Arabic countries need to be written in both languages, and most of the new and translated religious books are written in both languages by Muslim scholars. Similar to English, a large amount of unstructured documents are available on the net in Arabic and Amharic languages. However, IR tools and techniques are mostly English language oriented, and currently there are several attempts to develop IR tools for Arabic and Amharic language. Many of Internet users who are non-native Arabic speakers can read and understand Arabic documents but they feel uncomfortable to formulate queries in Arabic. This may be either because of their limited vocabulary in Arabic, or because of the possible miss-usage of Arabic words. Different attempts have been made to develop CLIR systems for Amharic-French [19] and Afan Oromo-English [3] languages. Nevertheless, CLIR system is not found for Amharic-Arabic language pair.

Development of standard corpus and tools is very essential in order to test the performance of the newly developed CLIR system [20]..

The aim of this research work is to develop a prototype of dictionary based Amharic-Arabic CLIR system that enables Amharic and Arabic language users to retrieve both language documents and to examine the ability of the proposed system. We employee query translation strategy, which is more efficient than document translation strategy, because the document translation strategy require overhead cost of translating all documents, especially when new documents are added frequently and not all of the documents are of interest to the users [21].

The remainder of this paper is organized as follows; the review of related works is presented in Section 2 and the proposed CLIR method in Section 3. Section 4 gives the experimental setup and the results and the paper conclude in Section 5.

2. RELATED WORKS

Several researchers have studied CLIR works related to different language pairs. However, less work is reported on Amharic and Arabic languages paired with other languages. Some of the prominent works are discussed below

Argaw Atelach Alemu, et.al [19], present a dictionary based approach to translate the Amharic queries into French Bags-of-words in the Amharic-French bilingual track at CLEF 2005 using the search engines: SICS and Lucene. Non-content bearing words were removed both before and after the dictionary lookup. TF/IDF values supplemented by a heuristic function was used to remove the stop words from the Amharic queries and two French stop words lists were used to remove stop words from French translations. From the experiments, they found that the SICS search engine performed better than Lucene. Aljlayl et.al [1], empirically evaluated the use of an MT-based approach for query translation in an Arabic-English CLIR system using TREC-7 and TREC-9 topics and collections. The effect of query length on the performance of MT is also investigated to explore how much context is actually required for successful MT processing. A well-formed source query makes the MT system able to provide its best accuracy. Tesfaye Fasika [20], employed a corpus based approach which makes use of phrasal query translation for Amharic-English CLIR. The result of the experimentation is a recall value of 24.8% for translated Amharic queries, 46.3% for Amharic queries and 43.6% for the baseline English queries. Nigussie Eyob [7], have developed a corpus based Afaan Oromo–Amharic CLIR system to enable Afaan Oromo speakers to retrieve Amharic information using Afaan Oromo queries. Documents including news articles, bible, legal documents and proclamations from customs authority were used as parallel corpus. Two experiments were conducted, by allowing only one possible translation to each Afaan Oromo query term and by allowing all possible translations. The first experiment returned a maximum average precision of 81% and 45% for monolingual (Afaan Oromo) queries and bilingual (translated Amharic) queries run respectively. The second experiment showed better result of recall and precision than the first experiment, which is 60% for the bilingual query run, and the result for the monolingual query run remained the same.

Mequannint et al. [22], designed a model for an Amharic-English Search Engine and developed a bilingual Web search engine based on the model that enables Web users for finding the information they need in Amharic and English languages. They have identified different language dependent query pre-processing components for query translation and developed a bidirectional dictionary-based translation system, which incorporates a transliteration component to handle proper names, which are often missing in bilingual lexicons. They used an Amharic search engine and an open source English search engine (Nutch) for Web document crawling, indexing,

searching, ranking and retrieving. The experimental results showed that the Amharic-English Cross-Lingual Retrieval engine performed 74.12% of its corresponding English monolingual retrieval engine and the English-Amharic Cross-Lingual Retrieval engine performed 78.82% of its corresponding Amharic monolingual retrieval engine.

In CLIR, the semantic level of words is crucial. Solving the problem of word sense disambiguation will enhance the effectiveness of CLIR systems. Andres Duque et al [23], studied to choose the best dictionary for Cross Lingual Word Sense Disambiguation (CLWSD). They applied the comparison between different dictionaries in two different frameworks; analysing the potential results of an ideal system using those dictionaries and considering the particular unsupervised CLWSD system Co-occurrence Graph, then analyse the results obtained when using different bilingual dictionaries providing the potential translations. They also developed hybrid system by combining the results provided by a probabilistic dictionary, and those obtained with a Most Frequent Sense (MFS) approach. They have focused on only on English- Spanish cross-lingual disambiguation. The hybrid approach outperforms the results obtained by other unsupervised systems.

As Arabic is a relatively widely researched Semitic language and has a number of common properties that share with Amharic, some of the computational linguistic research [1],[19],[24], conducted on Amharic and Arabic languages nowadays recommended customizing and using the tools developed for these languages. While the above researchers has attempted to develop and evaluate Amharic and Arabic paired languages with other languages separately, no research has these two languages paired together.

3. METHODOLOGY

In this work, an attempt has been made to design a dictionary based Amharic-Arabic CLIR system, which has indexing and searching tasks. Inverted file indexing structure is used to organize documents to speed up searching. The probabilistic model that attempts to simulate the uncertainty nature of an IR system guides the searching process. Amharic and Arabic documents are pre-processed separately by performing tokenization, normalization, stop word removal, punctuation removal and stemming. Figure 3.1 shows the general architecture of the system, which is adopted from C. Peters et al [25]. Bi-lingual dictionary, which includes the list of Amharic and Arabic translated words is constructed manually and is used to translate Amharic queries to Arabic queries.

Binary independent probabilistic information retrieval model is adopted to search the relevant documents from Amharic-Arabic parallel corpus. Probabilistic information retrieval is the estimation of the probability of relevance that a document di will be judged relevant by the user with respect to query q, which is expressed as, P(R|q, di), where, R is the set of relevant documents. Typically, in probabilistic model, based on the query the documents are divided into relevant and irrelevant documents [26]. However, the probability of any document is relevant or irrelevant with respect to users query is initially unknown. Therefore, the probabilistic model needs to guess the relevance at the beginning of search process. The user then observes the first retrieved documents and gives feedback for the system by selecting relevant documents as relevant and irrelevant documents as irrelevant. By collecting relevance feedback data from a few documents, the model can then be applied to estimate the probability of relevance for the remaining documents in the collection. This process is applied iteratively to improve the

performance of the system to retrieve more and more relevant documents, which satisfies the users need.

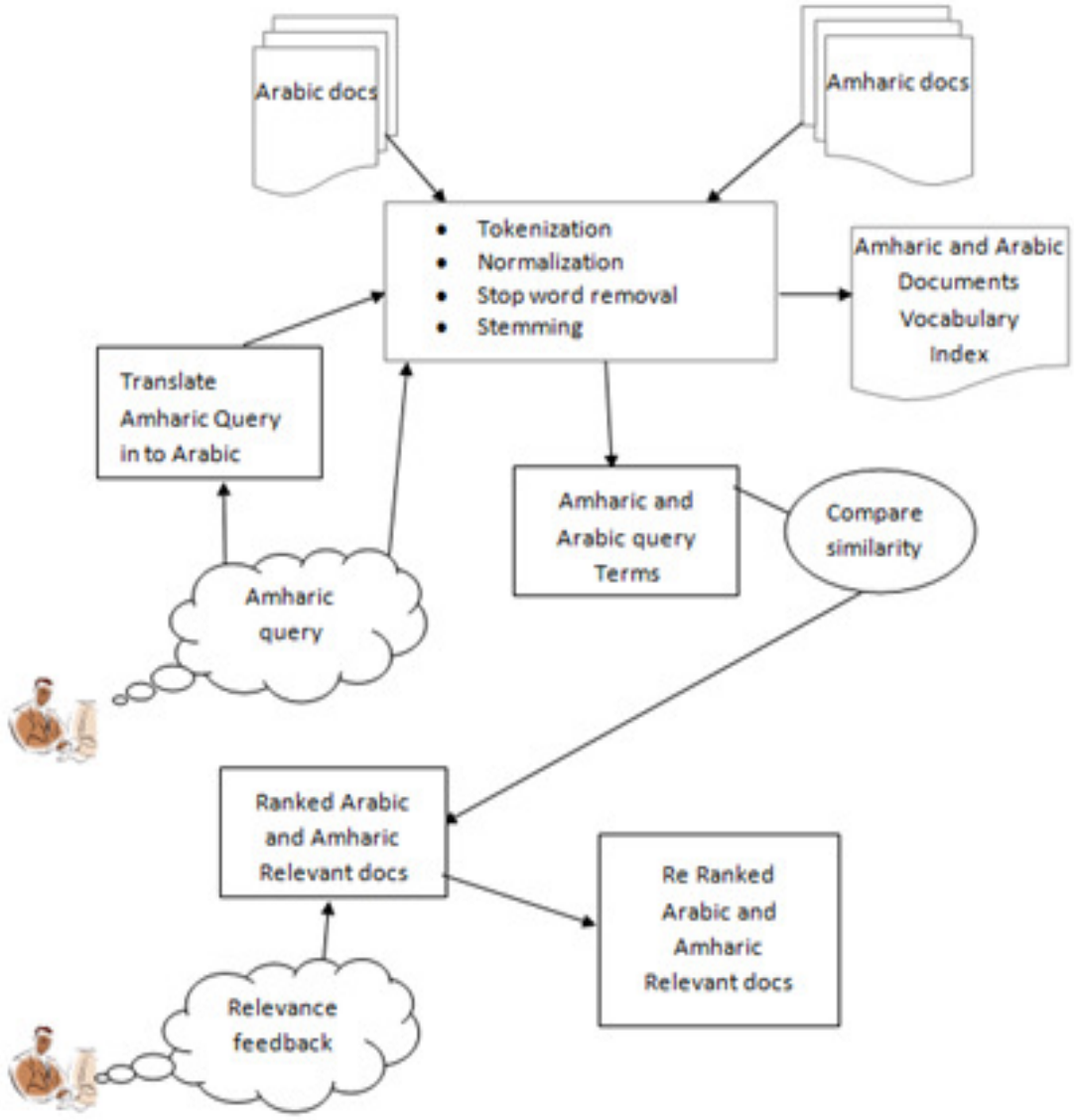

Figure 3.1 Dictionary based Amharic-Arabic CLIR system architecture

The assumptions made for the uncertainty nature of probability model are;

- $p(k_i|R)$ is constant for all index terms k (usually, its equal to 0.5)
- The distribution of index terms among the non-relevant documents can be approximated by the distribution of index terms among all the documents in the collection.

These two assumptions will give;

$$P(k_i|R) = 0.5 \text{ and } p(k_i|R) = \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right) \dots\dots\dots\dots\dots\dots\dots\dots\dots (1)$$

where, N is the total number of documents in the collection and n_i is the number of documents which contain the index term k_i.

4. EXPERIMENTATION AND EVALUATION

The Holy Quran available through Tanzile Quran navigator website [27] includes 114 chapters, each containing a minimum of 3 to a maximum of 286 verses in Arabic Amharic languages. In this work, subject to the availability of the number of verses, we have downloaded upto 10 verses from each chapter in Arabic and the corresponding verses in Amharic.

Even though complete evaluation process requires the evaluation of both system effectiveness and efficiency, only effectiveness of IR system is taken into consideration to determine the performance of the system for the translated queries. Precision and recall are used to measure the effectiveness of the IR system designed.

We used Amharic queries for the retrieval of documents both in Arabic and Amharic languages. In addition to retrieving Amharic documents, the Amharic query is translated into Arabic for retrieving Arabic documents. We used 14 simple queries to test the performance of the system and the results obtained are shown in Table 4.1. The performance of the system on Arabic relevant retrieved documents is much better than that of Amharic documents (i.e., 83.89% precision for Amharic against 52.02% precision for Arabic).

When the system is tested by giving queries that has Out of Vocabulary words in the dictionary, its precision is decreased and recall is increased specially for Arabic documents. For example, if we add a word "ለኾነው" (to become) which is not translated correctly or appeared in the dictionary for the first query "የፍርዱ ቀን ባለቤት ለኾነው" (Financed you day of the debt) the word "ለኾነው" (to become) is directly used for searching. Therefore, the number of Amharic non relevant documents increased by highly decreasing the performance of the system. The main hindrance of the system performance is incorrect translation due to unnormalized Arabic words specifically diacritics for mapped with the dictionary words, system that cannot be.

Table 4.1 Performance of the proposed system

Query No	Query in Amharic	Query translated to English	Amharic Documents		Arabic Documents	
			Precision %	Recall %	Precision %	Recall %
1	የፍርዱቀን ባለቤት ለኽ ው·	Financed you day of the debt	8.16	100	33.33	50
2	ፊቱን አፀፋገ ዞረም	Frowned watawallaY◌ that came him aal'aae'maY◌	100	33.33	100	33.33
3	ለእርስዎምአንደምቆጠየ ለወም	Does not cherish for his one refrained	25	100	80	80
4	ለጌታህ ስገድበስመለዋም	So your god arrives for waaanHar	20.83	100	100	20
5	ሰውበክሳራ ውስጥ ነ ው·	That the human is turning me lost	35.71	100	95.24	100
6	ሰምኮንም ፉፍሮት እየ ኽ· ሲገቡ ባየህ ጊዜ	The people saw debt of Allah enter in regiments	18.51	83.33	100	16.67
7	ቁረይሽን ለማለመጅ ባለገዘሆኖቹን አጠፉ·	For agreement of Quraish	100	50	100	50
8	ከፊጠረወፍጠፎ ሀሉ·ከፋሪት	From evil what created	70	71.43	32.56	100
9	የማ·ተግት ሲያጣ ከፋሪ አጠቢቃለሁ·	From evil of the delusion aalxannaaasi	50	60	100	40
10	ማ·ቱምአንፊ·ተሸካማሽትቸን	His authorities support of the wood	66.67	100	100	50
11	መጎንት·የ ኽኑን ምፍት·ላከ	Allah sent them flew 'aabaaabiyla	33.33	100	100	50
12	አናንተም እኔ የምገዛወን ተገ ፎች አይደላችሁም	Nor you is worshipers what worships	16.67	75	50	25
13	የዕቃ ትወስጥንም የማ·ለከሉ ለኽ·ት ወሳጥው·	The ream prevent	100	50	100	50
14	ከዋክብቱም በረገፉ· ተራራምነም በተነ ዱ·	If the stars aankadarat and if the mountainswalked	83.33	100	83.33	100
Weighted Average Precision and Recall (WAPR)			52.02	80.22	83.89	54.64

5. CONCLUSION

Multilingual information is required for the countries that have multiple languages and it is vital as the users of the internet throughout the world are ever increasing. We have developed a prototype of dictionary based Amharic-Arabic CLIR system that enables Amharic and Arabic language users to retrieve both language documents and to examine the ability of the proposed system. The effectiveness of our proposed system was evaluated and the performance of the system on Arabic relevant retrieved documents was much better than that of Amharic documents.

The main challenges with dictionary-based CLIR are untranslatable words due to the limitation of Amharic Arabic general dictionary, the processing of inflected words, Phrase identification and translation, and lexical ambiguity in Amharic and Arabic language.

Even if this research has a vital significance in retrieving the required information from Amharic-Arabic document, some issues need to be further investigated to develop efficient and effective CLIR system. This approach requires an exhaustive and detailed list of mapping of concepts in both languages, which is very difficult to build.

REFERENCES

[1] M. Aljlayl, O. Frieder, and D. Grossman, "On Arabic-English cross-language information retrieval: A machine translation approach," in Information Technology: Coding and Computing, 2002. Proceedings. International Conference on, 2002, pp. 2–7.

[2] K. Sourabh, "An Extensive Literature Review on CLIR and MT activities in India," Int. J. Sci. Eng. Res., 2013.

[3] D. Bekele, "Afaan Oromo Oromo-English Cross-Lingual Information Retrieval (Clir)," AAU, 2011.

[4] D. Kelly, "Methods for evaluating interactive information retrieval systems with users," Found. Trends Inf. Retr., vol. 3, no. 1—2, pp. 1–224, 2009.

[5] J. Cardeñosa, C. Gallardo, and A. Toni, "Multilingual Cross Language Information Retrieval A new approach."

[6] M. Abusalah, J. Tait, and M. Oakes, "Literature Review of Cross Language Information Retrieval," Comput. Hum., pp. 175–177, 2005.

[7] E. Nigussie, "Afaan Oromo--Amharic Cross Lingual Information Retrieval," AAU, 2013.

[8] T. Hedlund, "Dictionary-based cross-language information retrieval: principles, system design and evaluation," in SIGIR Forum, 2004, vol. 38, no. 1, p. 76.

[9] M. R. Warrier and M. S. S. Govilkar, "A SURVEY ON VARIOUS CLIR TECHNIQUES."

[10] D. Zhou, M. Truran, T. Brailsford, V. Wade, and H. Ashman, "Translation techniques in cross-language information retrieval," ACM Comput. Surv., vol. 45, no. 1, p. 1, 2012.

[11] G.-A. Levow, D. W. Oard, and P. Resnik, "Dictionary-based techniques for cross-language information retrieval," Inf. Process. Manag., vol. 41, no. 3, pp. 523–547, 2005.

[12] A. D. Rubin, "The Subgrouping of the Semitic Languages," Linguist. Lang. Compass, vol. 2, no. 1, pp. 79–102, 2008.

[13] S. ABEBAYEHU, "Amharic-English Script Identification in Real-Life Document Images," aau, 2012.

[14] B. Ayalew, "The submorphemic structure of Amharic: toward a phonosemantic analysis," University of Illinois at Urbana-Champaign, 2013.

[15] R. Tsarfaty, "Syntax and Parsing of Semitic Languages," in Natural Language Processing of Semitic Languages, Springer, 2014, pp. 67–128.

[16] H. Ishkewy, H. Harb, and H. Farahat, "Azhary: An arabic lexical ontology," arXiv Prepr. arXiv1411.1999, 2014.

[17] T. Hailemeskel, "Amharic Text Retrieval: An Experiment Using Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD)," M. Sc. Thesis, Addis Ababa University, Addis Ababa, 2003.

[18] F. Ahmed and A. Nurnberger, "Arabic/English word translation disambiguation approach based on na{"\i}ve Bayesian classifier," in Computer Science and Information Technology, 2008. IMCSIT 2008. International Multiconference on, 2008, pp. 331–338.

[19] A. A. Argaw, L. Asker, J. Karlgren, M. Sahlgren, and R. Cöster, "Dictionary-based Amharic-French information retrieval," CEUR Workshop Proc., vol. 1171, 2005.

[20] F. Tesfaye, "Phrasal Translation for Amharic English Cross Language Information Retrieval (Clir)," AAU, 2010.

[21] M. Adriani, "Using statistical term similarity for sense disambiguation in cross-language information retrieval," Inf. Retr. Boston., vol. 2, no. 1, pp. 71–82, 2000.

[22] M. Munye and S. Atnafu, "Amharic-English bilingual web search engine," in Proceedings of the International Conference on Management of Emergent Digital EcoSystems, 2012, pp. 32–39.

[23] A. Duque, J. Martinez-Romo, and L. Araujo, "Choosing the best dictionary for Cross-Lingual Word Sense Disambiguation," Knowledge-Based Syst., vol. 81, pp. 65–75, 2015.

[24] S. A. L. S. F. Adafre, "Machine Translation for Amharic: Where we are," Strateg. Dev. Mach. Transl. Minor. Lang., p. 47.

[25] C. Peters, M. Braschler, and P. Clough, Multilingual information retrieval: From research to practice. Springer Science & Business Media, 2012.

[26] F. Dahak, M. Boughanem, and A. Balla, "A probabilistic model to exploit user expectations in XML information retrieval," Inf. Process. Manag., 2016.

[27] "http://tanzil.net/#trans/am.sadiq." .

AUTHORS

Ibrahim Gashaw Kassa, is a Ph.D. candidate at Mangalore University Karnataka State, India since 2016. He graduated in 2006 in Information System from Addis Ababa University, Ethiopia. In 2014, he obtained his master's degree in Information Technology from University of Gondar, Ethiopia., he serves as a lecturer at University of Gondar from 2009 to May 2016. His research interest is in Cross Language Information Retrieval.

Dr. H L Shashirekha is an Associate Professor in the Department of Computer Science, Mangalore University, Mangalore, Karnataka State, India. She completed her M.Sc. in Computer Science in 1992 and Ph.D. in 2010 from University of Mysore. She is a member of Board of Studies and Board of Examiners (PG) in Computer Science, Mangalore University. She has presented several papers in International Conferences and published several papers in International Journals and Conference Proceedings. Her area of research includes Text Mining and Natural Language Processing.

AUTHOR INDEX