Online ISSN : 0975-4172
Print ISSN : 0975-4350
DOI : 10.17406/GJCST

Global Journal

of Computer Science and Technology: C

Software & Data Engineering

Highlights

- Sentiment Analysis
- Review of Feature Selection
- Security in Data Mining
- Evaluation of Features Extraction

Discovering Thoughts, Inventing Future

Volume 16 Issue 5 Version 0.1

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Global Journal of Computer Science and Technology: C
Software & Data Engineering

Volume 16 Issue 5 (Ver. 1.0)

Open Association of Research Society

The opinions and statements made in this book are those of the authors concerned. Ultraculture has not verified and neither confirms nor denies any of the foregoing and no warranty or fitness is implied.

Engage with the contents herein at your own risk.

The use of this journal, and the terms and conditions for our providing information, is governed by our Disclaimer, Terms and Conditions and Privacy Policy given on our website http://globaljournals.us/terms-and-condition/menu-id-1463/

By referring / using / reading / any type of association / referencing this journal, this signifies and you acknowledge that you have read them and that you accept and will be bound by the terms thereof.

All information, journals, this journal, activities undertaken, materials, services and our website, terms and conditions, privacy policy, and this journal is subject to change anytime without any prior notice.

Incorporation No.: 0423089
License No.: 42125/022010/1186
Registration No.: 430374
Import-Export Code: 1109007027
Employer Identification Number (EIN):
USA Tax ID: 98-0673427

Global Journals Inc.

(A Delaware USA Incorporation with “Good Standing”; ***Reg. Number: 0423089****)*

Sponsors: Open Association of Research Society
Open Scientific Standards

Publisher’s Headquarters office

Global Journals® Headquarters
945th Concord Streets,
Framingham Massachusetts Pin: 01701,
United States of America
USA Toll Free: +001-888-839-7392
USA Toll Free Fax: +001-888-839-7392

Offset Typesetting

Global Journals Incorporated
2nd, Lansdowne, Lansdowne Rd., Croydon-Surrey,
Pin: CR9 2ER, United Kingdom

Packaging & Continental Dispatching

Global Journals
E-3130 Sudama Nagar, Near Gopur Square,
Indore, M.P., Pin: 452009, India

Find a correspondence nodal officer near you

To find nodal officer of your country, please email us at *local@globaljournals.org*

eContacts

Press Inquiries: *press@globaljournals.org*
Investor Inquiries: *investors@globaljournals.org*
Technical Support: *technology@globaljournals.org*
Media & Releases: *media@globaljournals.org*

Pricing (Including by Air Parcel Charges):

For Authors:
22 USD (B/W) & 50 USD (Color)
Yearly Subscription (Personal & Institutional):
200 USD (B/W) & 250 USD (Color)

Global Journals Constitutional Editorial Board

~Integrated~

Contents of the Issue

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Feature Extraction and Duplicate Detection for Text Mining: A Survey

By Ramya R S, Venugopal K R, Iyengar S S & Patnaik L

University Visvesvaraya College of Engineering

Abstract- Text mining, also known as Intelligent Text Analysis is an important research area. It is very difficult to focus on the most appropriate information due to the high dimensionality of data. Feature Extraction is one of the important techniques in data reduction to discover the most important features. Proce- ssing massive amount of data stored in a unstructured form is a challenging task. Several pre-processing methods and algo- rithms are needed to extract useful features from huge amount of data. The survey covers different text summarization, classi- fication, clustering methods to discover useful features and also discovering query facets which are multiple groups of words or phrases that explain and summarize the content covered by a query thereby reducing time taken by the user.

Keywords: text feature extraction, text mining, query search, text classification.

GJCST-C Classification: C.2.1,C.2.4,H.2.8

Feature Extraction and Duplicate Detection for Text Mining: A Survey

Ramya R S $^{\alpha}$, Venugopal K R $^{\sigma}$, Iyengar S S $^{\rho}$ & Patnaik L M $^{\omega}$

Abstract- Text mining, also known as Intelligent Text Analysis is an important research area. It is very difficult to focus on the most appropriate information due to the high dimensionality of data. Feature Extraction is one of the important techniques in data reduction to discover the most important features. Processing massive amount of data stored in a unstructured form is a challenging task. Several pre-processing methods and algorithms are needed to extract useful features from huge amount of data. The survey covers different text summarization, classification, clustering methods to discover useful features and also discovering query facets which are multiple groups of words or phrases that explain and summarize the content covered by a query thereby reducing time taken by the user. Dealing with collection of text documents, it is also very important to filter out duplicate data. Once duplicates are deleted, it is recommended to replace the removed duplicates. Hence we also review the literature on duplicate detection and data fusion (remove and replace duplicates).The survey provides existing text mining techniques to extract relevant features, detect duplicates and to replace the duplicate data to get fine grained knowledge to the user.

Keywords: text feature extraction, text mining, query search, text classification.

I. Introduction

Society is increasingly becoming more digitized and as a result organisations are producing and storing vast amount of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Web based social applications like people connecting websites results in huge amount of unstructured text data. These huge data contains a lot of useful information. People hardly bother about the correctness of grammar while forming a sentence that may lead to lexical syntactical and semantic ambiguities. The ability of finding patterns from unstructured form of text data is a difficult task.

Data mining aims to discover previously unknown interrelations among apparently unrelated attributes of data sets by applying methods from several areas including machine learning, database systems, and statistics. Many researches have emphasized on different branches of data mining such as opinion mining, web mining, text mining. Text mining is one of the most important strategy involved in the phenomenon of knowledge discovery. It is a technique of selecting previously unknown, hidden, understandable, interesting knowledge or patterns which are not structured. The prime objective of text mining is to diminish the effort made by the users to obtain appropriate information from the collection of text sources [1].

Thus, our focus is on methods that extract useful patterns from texts in order to categorize or structure text collections. Generally, around 80 percent of company's information is saved in text documents. Hence text mining has a higher economic value than data mining. Current research in the area of text mining tackles problems of text representation, classification, clustering, information extraction or the search for and modelling of hidden patterns. Selection of characteristics, influence of domain knowledge and domain-specific procedures play an important role.

The text documents contain large scale terms, patterns and duplicate lists. Queries submitted by the user on web search are usually listed on top retrieved documents. Finding the best query facet and how to effectively use large scale patterns remains a hard problem in text mining. However, the traditional feature selection methods are not effective for selecting text features for solving relevance issue. These issues suggests that we need an efficient and effective methods to mine fine grained knowledge from the huge amount of text documents and helps the user to get information quickly about a user query without browsing tens of pages.

The paper provides a review of an innovative techniques for extracting and classifying terms and patterns. A user query is usually presented in list styles and repeated many times among top retrieved documents. To aggregate frequent lists within the top search results, various navigational techniques have been presented to mine query facets.

The Organisation of the paper is as follows: Section 1 introduces a detailed overview of text mining frameworks, application and benefits of text mining. Sections 2 and 3 reviews feature selection, feature extraction and techniques of pattern extraction. Section 4 discusses various text classification and clustering algorithms in text mining. Sections 5 and 6 introduce a detailed overview of discovering facets and fine grained knowledge. Section 7 reviews the duplicate detection in text documents. Section 8 contains the conclusions.

a) Text Mining Models

Text mining tasks consists of three steps: text preprocessing, text mining operations, text post processing. Text preprocessing includes data selection,

Author α σ ρ Ѡ: Department of Computer Science and Engineering University Visvesvaraya College of Engineering, Bangalore University, Bangalore 560 001. e-mail: rs.ramya.reddy@gmail.com

text categorization and feature extraction. Text mining operations are the core part of text mining that includes association rule discovery, text clustering and pattern discovery as shown in Figure1. Post processing tasks modifies the data after text mining operations are completed such as selecting, evaluating and visualization of knowledge. It consists of two components text filtering and knowledge cleansing. Many approaches [2] have been concerned of obtaining structured datasets called intermediate forms, on which techniques of data mining [3] are executed.

Text filtering translates collection of text documents into selected intermediate form(IF) which means Knowledge cleansing or discovering patterns. It can be structured or semistructured. Text mining methods like clustering, cla- ssification and feature extraction falls within document based IF. Pattern discovery and relationship of the obje- ct, associative discovery, visualization fall within object based documents.

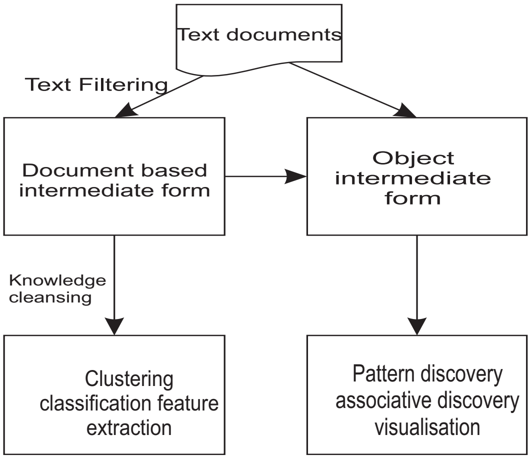

Text documents
Text Filtering
Document based intermediate form
Object intermediate form
Knowledge cleansing
Clustering classification feature extraction
Pattern discovery associative discovery visualisation

Figure 1: Text mining framework

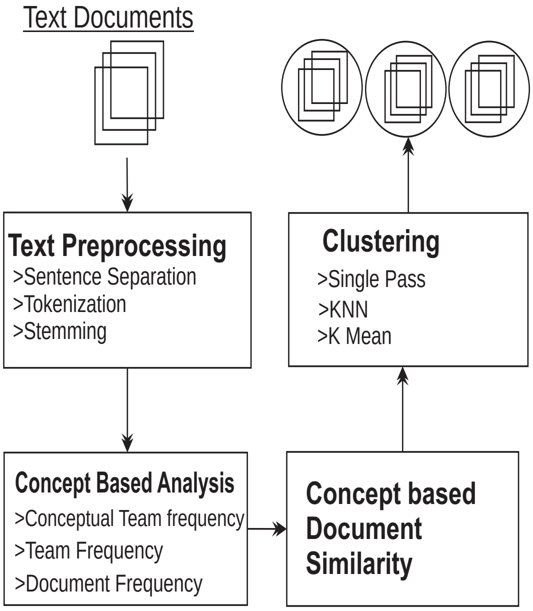


Figure 2: Concept based Mining Model System

When documents contains terms with same frequency. Two terms can be meaningful while the other term may be irrelevant. Inorder to discover the semantic of text, the mining model is introduced. Figure 2 represents a new mining model based on concepts. The model is proposed to analyse terms in a sentence from documents. The model contains group of concept analysis, they are sentence based concept analysis, document based concept analysis and corpus based similarity measure [4]. Similarity measure concept based analysis calculates the similarity between documents. The model effectively and efficiently finds matching concepts between documents, according to the meaning of their sentence.

b) Application and benefits of Text Mining

Text mining [5] [6] [7] has several applications in discovering hidden knowledge and it can be used in one of the three ways.

Clustering: The process of grouping similar kind of information is called clustering that results in finding interesting knowledge. The new discovered knowledge can be used by an industry for further development and helps in competing with their competitors.

Question Answering: For seperating and combining terms we use standard text searching techniques that use boolean operators. Sophisticated search in text mining executes the searching process in sentence or phrase level and verbal connection identification between various search terms, which is not possible in traditional search. The result obtained by sophisticated search can be used for providing specific information that can be influenced by an organization.

Concept linkage: The results obtained from sophisticated search are linked together to produce a new hypothesis. The linking of concepts is called concept linkage. Hence, new domain of knowledge can be generated by making use of concept linkage application.

Benefits of text mining are better collection development to resolve user needs, information retrieval, to resolve usability and system performance, data base evaluation, hypothesis development. Information professionals(IP) [8] are always in forefront for emerging technologies. Inorder to make their product and service better and more efficient, usually libraries and information use these IP. The trained information professionals manage both technical and semantic infrastructures which is very important in text mining. IP also manages content selection and formulation of search techniques and algorithms.

Akilan et al., [9] pesented the challenges and future directions in text mining. It is mandatory to function semantic analysis to capture objects relationship in the documents. Semantic analysis is computationally expensive and operates on few words per second as text mining consists of significant language component. An effective text refining method has to be developed to process multilingual text document. Trained knowledge specialists are neceessary to deal with products and application of current text mining tools. Automated mining operations is required which can be used by technical users. Domain Knowledge plays an important role in both at text refining stage and knowledge distillation and hence helps in improving the efficiency of text mining.

Sanchez et al., [10] presented Text knowledge mining (TKM) based deductive inference that is usually targeted on the feasible subset of texts which usually search for contradictions. The procedure obtains new knowledge making a union of intermediate forms of texts from accurate knowledge expressed in the text.

Dai et al., [11] introduced competitive intelligence analysis methods FFA (Five Faces Frame work) and SWOT with text mining technologies. The knowledge is extracted from the raw data while performing transforming process that enables the business enterprises to take decisions more reliably and easily. Mining Environment for Decisions (MinEDec) system is not evaluated in real business environments.

Hu et al., [12] presented a interesting task of automatically generating presentation slides for academic papers. Using a support vector regression method, importance scores of sentences in the academic papers is provided. Another method called Integer Linear Programming is used to generate well structured slides. The method provides the researchers to prepare draft slides which helps in final slides used for presentation. The approach does not focus on tables, graphs and figures in the academic papers.

c) Traffic based Event in Text Mining

Andrea et al., [13] [14] have proposed a real-time monitoring system for traffic event detection that fetches tweets, classifies and then notifies the users about traffic events. Tweets are fetched using some text mining techniques. It provides the class labels to each tweet that are related to a traffic event. If the event is due to an external cause, such as football match, procession and manifestation, the system also discriminate the traffic event. Final result shows it is capable of detecting traffic event but traffic condition notifications in real-time is not captured.

An efficient and scalable system from a set of microblogs/ tweets has been proposed to detect Events from Tweets (ET) [15] by considering their textual and temporal components. The main goal of proposed ET system is the efficient use of content similarity and appearance similarity among keywords and to cluster the related keywords. Hierarchical clustering technique is used to determine the events, which is based on common co-occurring features of keywords [16]. ET is evaluated on two different datasets from two different domains. The results show that it is possible to detect events of relevance efficiently. The use of semantic knowledge base like Yago is not incorporated.

Schulz et al., [17] proposed a machine learning algorithm which includes text classification and increasing the semantics of the microblog. It identifies the small scale incidents with high accuracy. It also precisely localizes microblogs in space and time which enables it to detect incidents in real time. The algorithm will not only give us information about the incident and in addition give us valuable information on previous unknown information about the incidents. It does not considers NLP techniques and large data.

ITS (Intelligent Transportation Systems) [18] recognizes the traffic panels and dig in information contained on them. Firstly, it applies white and blue color segmentation and then at some point of interest it derives descriptors. These images that can now be considered as sack of words and classified using Na¨ıve Bayes or SVM (state vector method). The kind of categorization where the images are classified based on visual appearance is new for traffic panel detection and it does not recognize multiframe integration.

II. Preprocessing in Text Mining

Text may be loosely organized without complete information in the documents and may also contain omitted information. The text has to be scanned attentively to determine the problems. If it is not scanned and scrutinised properly then it leads to poor accuracy on unstructured data and hence preprocessing is necessary.

Preprocessing guarantees successful implementation of text analysis, but may spend substantial processing time. Text processing can be done in two basic methods. a)Feature Selection b) Feature Extraction.

a) Feature Selection

Research in numerous fields like machine learning, data mining, computer vision, statistics and linked fields has led to diversity of feature selection approaches in supervised and unsupervised surroundings.

Feature Selection (FS) has an important role in data mining in categorization of text. The centralized idea of feature selection is the reduction of the dimension of the feature set by determining the features appropriately which enhances the efficiency and the performance. FS is a search process and categorized into forward search and backward search.

Mehdi et al., [19] [20] executed a innovative feature selection algorithm based on Ant Colony Optimization (ACO).

Without any prior knowledge of features, a minimal feature subset is determined by applying ACO [21]. The approach uses simple nearest neighbor classifier to show the effectiveness of ACO algorithm by reducing the computational cost and it outperforms information gain and chi methods. Complex classifiers and different kinds of datasets are not incorporated. Combining feature selection algorithm with other population-based feature selection algorithms are not considered.

Gasca et al., [22] proposed feature selection method based on Multilayer Perceptron (MLP). Under certain objective functions the approach determines and also corrects proper set of irrelevant set of attributes. It further computes the relative contributions for individual attribute in reference to the units that are to be output. For each output unit, contribution are sorted in the descending order. An objective function called prominance is computed for each attribute. Selecting the features from large document faces problem in unsupervised learning because of unnamed class labels.

Sivagaminathan et al., [23] [24] proposed a fixed size subset, an hybrid approach to solve feature subset selection problem in neural network pattern classifier. It considers both the individual performance and subset performance. Features are selected using the pheromone trail and value of heuristic by state transition rules. After selecting the feature, the global updating rule takes place to increment the features, which ultimately gives better classification performance without increase in the overall computational cost.
selection algorithms.

Table 1: shows comparison of feature

Sl.no.	Authors	Feature Selection (FS)	Algorithm	Advantages	Disadvantages
1.	Zhao *et al.*,(2016) [25]	Unsupervised	Gradient	Preserve similarity and discriminant information, high clustering performance is achieved	Supervised FS is not considered
2.	Xu *et al.*,(2016) [26]	-	Deep Learning	Performs better than traditional dimensional reduction method	Meta data information of tweet is not considered
3.	Wang *et al.*,(2015) [27]	Supervised and Unsupervised	Global redundancy Minimization	Features are more compact and discriminant,Superior performance without parameter	-

Ogura et al., [28] proposed an approach to reduce a feature dimension space which calculates the probability distribution for each term that deviates from poissons. These deviations from poissons are non significant for the documents that does not belong to category. Three measures are employed as a benchmark and by using two classifiers SVM and K-NN gives better performance than other conventional classifiers. Gini index proved to be better than chisquare, IG in terms of macro, micro average values of F1. These measures do not utilize the number of times the term occurs in a document. The computational complexity could not be to suppressed for other typical measures such as information gain and CHI.

Feature selection is measured based on words term and document frequency. Azam et al., [29] observes these frequencies for measuring FS. The metrics of Discriminative Power Measure (DPM) and GINI index (GINI) are incorporated and the term frequency based metric is useful for small feature set. The most important features returned by DPM and GINI tend to discover most of the available information at a faster rate, i.e. against lower number of features. The DPM and GINI are comparatively slower in covering document frequency information.

Yan et al., [30] presented a graph embedded framework for dimensionality reduction. The framework is also used as a tool and unifies many feature extraction methods. Feature is selected based on spectral graph theory and proposed framework unifys both supervised and unsupervised feature selection.

Zhao et al., [31] developed a framework for preserving feature selection similarity to handle redundant feature. A combined optimization formulation of sparse multiple output regression formulation is used for selecting similarity preserving features. The framework do not address existing kernel, metric learning methods and semi-supervise feature selection methods.

1) *Feature Selection based Graph Reconstruction:*A Major task in efficient data mining is Feature selection. Feature selection has a significant challenge in small labeled-sample problem. If data is unlabeled then it is large. If the label of data is extremely tiny, then supervised feature selection algorithms fail for want of sufficient information.

Zhao et al., [32] introduced graph regularized data construction to overcome the problems in feature selection. The approach achieves higher clustering performance in both unsupervised and supervised feature selection.

Linked social media crops enormous amount of unlabeled data. In the prevailing system, selecting features for unlabeled data is a difficult task due to the lack of label information. Tang et al., [33] proposed an unsupervised feature selection framework, LUFS(Linked Unsupervised Feature Selection), for related social media data to surpass the problem. The design builds a pseudo-class labels through social dimension extraction and spectral analysis. LUFS efficiently exploits association information but does not exploit link information. Computer vision and pattern recognition problems are the two main problems which have inherent manifold structure. A laplacian regularizer is included to smoothen the clustering process along with the scale factor. In text mining applications, several existing systems incorporate a NLP-basedtechniques which parse the text and promote the usage patterns that is used for mining and examination of the parse trees that are trivial and complex.

Mousavi et al., [34] have formulated a weighted graph depiction of text, called Text Graphs that further captures grammar which serve as semantic dealings between words that are in textual terms. The text based graphs incorporates such a framework called SemScape that creates parse trees for each sentence and uses two step pattern based procedure for facilitation of extraction from parse trees candidate terms and their parsable grammar.

Due to the absence of label information, it is hard to select the discriminative features in unsupervised learning. In the prevailing system, unsupervised feature selection algorithms frequently select the features that preserve the best data dissemination. Yang et al., [35] proposed a new approach that is L2, 1 -norm regularized Unsupervised Discriminative Feature Selection (UDFS). The algorithm chooses the most discriminative feature subset from the entire feature set in batch mode. UDFS outclasses the existing unsupervised feature selection algorithms and selects discriminative features for data representation. The performance is sensitive to the number of selected features and is data dependent.

Cai et al., [36] presented a novel algorithm, called Graph regularized Nonnegative Matrix Factorization (GNMF) [37], which explicitly considers the local invariance. In GNMF, the geometrical information of the data space is pre-arranged by building a nearest neighbor graph and gathering parts-based representation space in which two data points are adequately close to each other, if they are connected in the graph. GNMF models the data space as a sub manifold rooted in the ambient space and achieves more discriminating power than the ordinary NMF approach.

Fan et al., [38] suggested a principled vibrational framework for unsupervised feature selection using the non Gaussian data which is subjective to several applications that range from several diversified domains to disciplines. The vibrational frameworks provides a deterministic alternative for Bayesian approximation by the maximization of a lower bound on the marginal probability which has an advantage of computational efficiency.

2) Text summarization and Dataset: Several approaches have been developed till date for automatic summarization by identifying important topic from single document or clustered documents. Gupta et al., [39] describes a topic representation approach that captures the topic and frequency driven approach using word probability which gives reasonable performance and conceptual simplicity.

Negi et al., [40] developed a system that summarizes the information from a clump of documents. The proposed system constructs the identifiers that are useful for retrieving the important information from the given text. It achieves high accuracy but cannot calculate the relevance of the document.

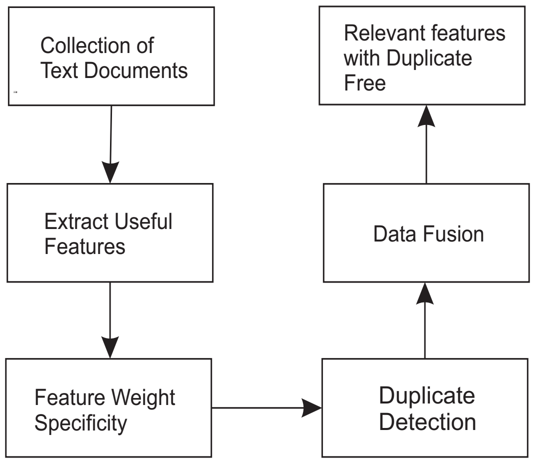
Debole et al., [41] initially explains the three phases in the life cycle of TC system like document indexing, classifier learning and classifier evaluation. All researches takes Reuters 21578 documents for TC experiments. Several researches have used Modapte split for testing. The three subsets used for the experiments are a set of ten categories with more number of positive training examples.

Xie et al., [42] proposed an approach to the acquisition of the semantic features within phrases from a single document that extracts document keyphrases. Keyphrase extraction method always performs better than TFIDF and KEA. Keyphrase extraction is a basic research in text mining and natural language processing. The method is developed on the concept of semantic relatedness where degrees between phrases are calculated by the cooccurrences between phrases in a given document and the same is presented as a relatedness graph. The approach is not domain specific and generalizes well on journal articles and is tested on news web pages.

To obtain any online information is an easy task. We log on to the world wide web and give simple keywords. However, it is not easy for the user to read the entire information provided. Hence text summarization is needed.

b) Feature Extraction

1) *Feature Mining for Text Mining:* Li et al.,[43] designed a new technique to discover patterns i.e., positive and negative in text document. Both relevant and irrelevant document contains useful features. Inorder to remove the noise, negative documents in the training set is used to improve the effectiveness of Pattern Taxonomy Model PTM. Two algorithms HLF mining and N revision was introduced. In HLF mining, it first finds positive features, discovers negative features and then composes the set of term. The offenders are selected by ranking the negative documents. The weights are initialized to the discovered terms of negative patterns. NRevision algorithms explains the terms weight based on their specificity and distribution in both positive and negative patterns.


Collection of Text Documents
Extract Useful Features
Feature Weight Specificity
Duplicate Detection
Data Fusion
Relevant features with Duplicate Free

Zhong et al., [44] has presented an effective pattern discovery technique which includes the process of pattern deploying and pattern evolving as shown in Table 2, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. The proposed model outperforms other pure data mining-based methods, the concept based models and term-based state-of the- art models, such as BM25 and SVM.

Li et al., [47] proposed two algorithms namely Fclustering and Wfeature to discover both positive and negative patterns in the text documents. The algorithm Fclustering classifies the terms into three categories general, positive, negative automatically without using parameters manually. After classifying the terms using Fclustering, Wfeature is executed to calculate the weights of the term. Wfeature is effective because the selected terms size is less than the average size of the documents. The proposed model is evaluated on RCV, Trec topics and Reuters 21578 dataset as shown in Table 2, the model performs much better than the term based method and pattern based method. The use of irrelevance feedback strategy is highly efficient for improving the overall performance of relevance feature discovery model.

Xu et al., [26] experimented on microblog dimensionality reduction- A deep learning approach. The approach aims at extracting useful information from large amount of textual data produced by microblogging services. The approach involves mapping of natural language texts into proper numerical representations which is a challenging issue. Two types of approaches namely modifying training data and modifying training objective of deep networks are presented to use micro-blog specific information. Meta-information contained in tweets like embedded hyperlinks is not explored.

Nguyen et al., [49] worked on review selection using Micro-reviews. The approach consists of two steps namely matching review sentences with micro reviews and selecting a few reviews which cover many reviews. A heuristic algorithm performs computionally fast and provides informative reviews.

2) *Feature Extraction for Classification:* Khadhim et al., [50] [21] developed two weighting methods TF -IDF and TF-IDF (Term Frequency/Inverse Document Frequency) global to reduce dimensionality of datasets because it is very difficult to process the original features i.e, thousands of features. Fuzzy c means clustering algorithm is used for feature extraction for classification.

3) *PCA and Random Projection RP:* Principal Component Analysis (PCA) is a simple technique used to explore and visualize the data easily. PCA extracts useful information from complicated data sets using non parametric method. It determines a lower dimension space by statistical method. Based on eigen value decomposition of the covariance matrix transformation matrix of PCA is calculated and thereby computation cost is more and it is also not suitable for very high dimensional data. The strength of PCA is that there are no parameters to fine tune and also no co-efficient is required to adjust.

Fradkin et al., [51] [52] reported a number of experiments by evaluating random projecton in supervised learning. Different datasets were tested to compare random projection and PCA using several machine learning methods. The results show PCA outperforms RP in efficiency for supervised learning. The results also shows that RP's are well suited to use with nearest neighbour and with SVM classifier and are less satisfactory with decision trees.

Table 2: Summary of The Feature Extraction

Sl.no.	Authors	Models Used for mining	Algorithm	Advantages	Disadvantages
1.	Chen *et al.,* (2016) [45]	-	Temporal pattern miner and Probabilistic Temporal Pattern Miner	Extracts interval based sequential patterns	Closed temporal patterns is not extracted
2.	Bartoli *et al.,* (2016) [46]	Genetic Programming	Regular Expressions	Handles large alphabets effectively	Snippets and faster learning is not considered
3.	Li *et al.,* (2015) [47]	Relevance Feature Discovery 1,2	WFeature Wclustering	Spec function is used for 3 categories and gives best performance,Automatically classifies the terms using Clustering method	Information is not utilized from non relevant documents
4.	Song *et al.,* (2013) [48]	-	Fast clustering based feature selection algorithm	Feature space dimensionality is reduced	Different types of co-relation measure are not explored
5.	Zhong *et al.,* (2012) [44]	Pattern Taxonomy Model	D pattern mining,IP Evolving	PTM is better than pattern based, concept based and term based models	-

III. Pattern Extraction

Patterns which are close to their super patterns that appears in the same paragragh are termed closed relation and needs to be eliminated. The shorter pattern is not considered since it is meaningless while the longer pattern is more meaningful and hence these are significant patterns in the pattern taxonomy.

Sequential Pattern (SP) mining algorithm is used to perform the extraction of frequent sequential pattern. It evaluates pruning and after pruning the final obtained patterns are added to next step of an algorithm. SP mining makes use of projected database method. The advantage of SP mining is that it deals with many sequences simultaneously whereas other techniques can handle one sequence at a time. The interesting information from negative or unlabeled documents are not extracted in this technique.

Abonem et al., [53] presented text mining framework that discovers knowledge by preprocessing the data. Usually text in the documents contains words, special characters and structural information and hence special characters is replaced by symbols. It mainly focuses on refining the uninterested patterns and thus fitering decreases the time and size of search space needed for the discovery phase. It is more efficient when large collection of documents are considered. Post-processing involves pruning, organizing and ordering of the results. The rule of each document is to find a set of characteristics phrases and keywords i.e., length, tightness and mutual confidence. The ranking of the rules within a document is measured by calculating a weight for each rule.

a) *Mining Closed Pattern*

Mining entire set of frequent subsequence for every long pattern generates uncontrollable number of frequent subsequence which are expensive in space and time. Yan et al., [54] proposed a solution for mining only frequent closed subsequence through an algorithm Clospan-Closed Sequential Pattern Mining. Clospan efficiently mines frequent closed sequences in large data sets with low minimum support but does not take advantage of search space pruning property.

Gomariz et al., [55] presented CSpan algorithm for mining closed sequential patterns which mines closed sequential patterns early by using pruning method called occurence checking. CSpan outperforms clospan and claspalgorithm.

Pei et al., [56] proposed CLOSET algorithm for discovering frequent closed itemsets. Three techniques extension of FP growth are developed to reduce the search space and to recognize the frequent closed itemsets. New strategies are constructed based on projection mechanism that makethe mining task scalable and efficient for massive database. CLOSET is faster than earlier methods for finding frequent closed itemsets.

b) *Mining Sequential Pattern*

To delimit the search and to increase the subsequence fragments Han et al., [57] proposed Freespan Frequent Pattern Projected sequential pattern Mining. Freespan fuses the mining of frequent sequence with that of frequent patterns and adopts projected sequence databases. Freespan runs quicker than the Apriori based GSP algorithm. Freespan is highly scalable and processing efficient in mining complete set of patterns. Freespan causes page thrashing as it requires extra memory. With extensive applications in data mining, mining sequential pattern encounters problems with a usage of very large database.

Pei et al., [58] proposed a sequential pattern mining method called Prefix Span(Prefix Projected sequential pattern mining). The complete set of patterns is extracted by reducing the generation of candidate subsequence. Further prefix projection largely reduces projected database size and greatly improves efficiency as shown in Table 3. Making use of RE(Regular Expression) [59] as a flexile constraint SPIRIT algorithm was proposed by Garofalakis et al., [60] for mining patterns that are sequential. A family of four algorithms is executed for forwarding a stronger relaxation of RE. Candidate sequence containing elements are pruned that do not appear in RE than its predecessor in the pattern mining loop.

The degree to which RE constraints are enforced to prune the search space of patterns are the main distinctive factor. The results on the real life data shows RE's adaptability as a user level tool for focussing on interesting patterns.

Jian et al., developed a new framework called Pattern Growth [PG]. PG is based on prefix monotonic property. Every monotonic and anti monotonic regular expression constraints are preprocessed and are pushed into a PG-based mining algorithm. PG adopts and also handles regular expression constraints which is difficult to explore using Apriori based method like SPIRIT. The candidate generation and test framework adopted by PG is less expensive and efficient in pushing many constraints than SPIRIT method. During Prefix growth various irrelevant sequence can be excluded in the huge dataset. Accordingly, projected database quickly shrinks. While PG outperforms SPIRIT, interesting constraints specific to complex structure mining is not be explored.

To filter the discovered patterns, Li et al., [43] [61] proposed an effective pattern discovery technique that deploys and evolves patterns to refine the discovered patterns. Using these discovered patterns, the relevant information can be determined inorder to improve the effectiveness. All frequent short patterns and long patterns are not useful and also long patterns with high specificity suffers from the low problem frequency. The problem of low frequency and misinterpretation for text mining can be solved by employing pattern deploying strategies.

Rather than using individual words, some researches used phrases to discover relevant patterns from documents collection. Hence there is a small improvement in the effectiveness of text mining becauses phrases based methods have consistency of assignment and document frequency for terms to be low. Inje et al., [62] used a pattern based taxonomy(is-a) relation to represent document rather than using single word. The computation cost is reduced by pruning unwanted patterns and hence improves the effectiveness of system.

Bayardo et al., [63] evaluated Max miner algorithm inorder to mine maximal frequent itemsets from large databases. Max- Miner reduces the space of itemsets considered through superset-frequency based pruning. There is a performance improvements over Apriori-like algorithms when frequent itemsets are long and more modest though still substantial improvements when frequent itemsets are short. Completeness at low supports on complex datasets is not achieved.

Jan et al., [64] [65] proposed propositionalization and classification that employs long first order frequent patterns for text mining. The Framework solves three text mining tasks such as information extraction, morphological disambiguation and context sensitive text correction. Propositionalization approach outperforms CBA by using frequent patterns as features. The performance of CBA classifiers greatly depends on number of class association rules and threshold values given by the user. The proposed framework shows that the distributed computation can improve performance of both method since large sample of data and a larger number offeatures are extracted.

Seno et al., [66] proposed an algorithm SLP miner that finds all sequential patterns. It performs effectively satisfying length decreasing support constraint and increases in average length of the sequence. It is expensive as pruning is not considered in this work.

Nizar et al., [67] demonstrates a taxonomy of sequential pattern mining techniques. Reducing the search space can be done by strongly minimizing the support count. Domain knowledge, distributed sequence are not considered in the mining process.

c) *Mining Frequent Sequences*

To extract sequential patterns, various algorithms have been executed by making continuously repeated scans of database and making use of hash structure.

Zaki et al., [68] presented a new novel algorithm SPADE for discovering sequential patterns at a high speed. SPADE decomposes the parent class into small subclasses. These sub problems are executed without depending on other subproblems in main memory by lattice approach. The lattice approach needs only one scan when having some pre-processed data. They also process depth first search and breadth first search for frequent sequence enumeration within each sublattice. By using these search strategies SPADE minimizes the computational costs and I/O costs by reducing number of database scans. It provides pruning strategies to identify the interesting patterns and prune out irrelevant patterns.

BFS outperforms DFS by having more information available for pruning while constructing a set of three sequence, two sequence. BFS require more main memory than DFS. BFS checks the track of idlists for all the classes, while DFS needs to preserve intermediate id lists for two consecutive classes along a specific path.

Han et al., [69] proposed a FP(frequent pattern tree) structure where the complete set of frequent patterns can be extracted by pattern fragment growth. Three techniques are used to achieve mining efficiency compression of the database, (i) FP tree avoids expensive repeatedly scanning database (ii) FP tree prevents generation of large number of candidates sets and uses divide and conquer method which breaks the mining task into a set of tasks that lowers search space. FP growth method [70] is efficient and also scalable for extracting both long and short frequent patterns and it is faster than Apriori algorithm.

Zhang et al., [71] executed CFP Constrained Frequent Pattern algorithm to improve the efficiency of association rule mining. The algorithm is incorporated in an interrelation analysis model for celestial spectra data. The module extracts correlation among the celestial spectra data characteristics. The model do not support for different application domain.

d) *Mining Frequent itemsets using Map Reduce*

Database Management System have evolved over the last four decades and now functionally rich. Operating and managing very large amount of business data is a challenging task. MapReduce [72] [73] is a framework that process and manages a very large datasets in a distributed clusters efficiently and achieves parallelism.

Xun et al., [74] [75] executed Fidoop algorithm using mapreduce model. Fidoop algorithm uses frequent itemset with different lengths to improve workload balance metric across clusters. Fidoop handles very high dimensional data efficiently but do not work on heterogeneous clusters for mining frequent itemsets.

Wang et al., [76] proposed (FIMMR) Frequent Itemset Mining Mapreduce Framework algorithm. The algorithm initially extracts lower frequent itemset, applies pruning technique and later mines global frequrnt itemset. The speedup of algorithm is satisfactory under low minimum support threshold.

Ramakrishnudu et al., [77] finds infrequent itemset from huge data using mapreduce framework. The efficiency of framework increases as the size of the data is increased. The framework produces few intermediate items during the process.

Ozkural et al., [78] extracts frequent item set by partitioning the graph by a vertex separator. The separator mines the item distribution independently. Parallel frequent itemset algorithm replicates the items that co-relate with the separator. The algorithm minimizes redundancy and load balancing is achieved. Relationship among a very large number of items for real world database is not incorporated.

e) *Relevance Feedback Documents*

Xu et al., [79] presented a Expectation Maximization(EM) algorithm for relevance feedback inoverlaps in feedback documents. Based on dirichlet compound multinominal(DCM) distribution, EM includes a background collection model reduction, by the methodology of deterministic annealing and query based regularization.

Several Queries which do not contain any relevance feedback needs improvisations by combining pseudo relevance feedback and relevance feedback using a hybrid feed-back paradigm. Instead of using static regularization, the authors adjust the regularization parameter based on the percentage of relevant feedback documents [80]. Further, the design formulates the space for a much newer document progressively. The weighted relevance is computed for an experimental design which further exploits the top retrieved documents by adjusting the selection scheme. The relevance score algorithms need to be validated on several TREC datasets.

Cao et al., [81] re-examined the assumption of most frequent terms in the false feedback documents that are useful and prove that it does not hold in reality. Distinguishing good and bad expansion terms cannot be done in the feedback documents. The difference of term distribution between feedback documents and whole document collection is exploited through the mixture model indicates that good and bad expansion terms may have similar distributions that fails to distinguish. Experiments are conducted to see that each query can keep only the good expansion terms. The new query model integrates the good terms, while classification of term is done to improve the effectiveness of retrieval. In a final query model, the classification score is used to enhance the weight of good terms. Selecting expansion terms are significantly better than traditional expansion terms by evaluating on three TREC datasets. Selection of terms has to be done carefully.

Pak et al., [82] proposed a automatic query expansion algorithm which incorporates a incremental blind approach to choose feedback documents from the top retrieved lists and further finds the terms by aggregating the scores from each feedback document. The algorithm performs significantly better on large documents.

Algarni et al., [83] proposed the adaptive relevance feature discovery(ARFD). Using a sliding window over positive and negative feedback, that ARFD updates the systems knowledge. The system provides a training documents where specific features are discovered. Various methods have been used to merge and revise the weight of the feature in a vector space. Documents are selected based on two categories. The first category is that user provide the interested topic information and the second category is that the user changed the interest topic.

IV. Text Classification and Clustering

Text categorization [84] is a significant issue in text mining. In general, the documents contains large texts and hence it is necessary to classify them into specific classes. Text categorization can be broadly classified into supervised and unsupervised classification. Classifying documents manually is very costly and time consuming task. Hence it is necessary to construct automaic text classifiers using pre-classified sample documents whose time efficiency and accuracy is much better than manual text classification.

Computer programs often treat the document as a sack of words. The main characteristics of text categorization is feature space having high dimensionality. Even for moderate sized text documents, the feature space consists of hundreds and thousands of terms.

Sebastiani et al., [85] reviews the standard approaches that comes under machine learning paradigm for text categorization. The approach also describes the prob- lem faced while document representation constructing classifiers and evaluation of constructed classifier. The experimental study shows comparisons among different classifiers on different versions of reutor dataset. Text categorization is a good benchmark for clarrifying whether a given learning technique can scale up to substantial sizes.

Irfan et al., [86] reviews different pre-processing techniques in text mining to extract various textual patterns from the social networking sites. To explore the unstructured text available from social web, the basic approaches of text mining like classification and clustering are provided.

Wu et al., [87] presents a technique consisting of three preprocessing stages to recognize the text region of huge size and contrast data. A Segmentation algorithm cannot identify the changes that happen both in color and illumination of character in a document image. The technique followsextracting the grayscale image such as from the book cover, magazine RGB plane associated with weighted valve. A multilevel thresholding process is done on each grayscale image independently to identify text region. A recursive filter is executed to interpret which connects components is textual components. An approach to determine score is considered to findout the probabilistic text region of resultant images. If the text region has maximum score, then it is classified as textual component.

V. Discovering Facets for Queries from Search Result

Facets means a phrase or a word. A query facet is a set of items which summarize an important aspect of a query. Dou et al., [88] [89] [90] explores solution of searching the set of facets for a user query. A system called Query Discovery (QD) miner is proposed to mine facets automatically. Expermiments are conducted for 100's of queries and results shows the effectiveness of the system as shown in the table 5. It provides interesting knowledge about a query and however improves searching for the users in different ways. The problem of generating query suggestions based on query facets is not considered that might help users find a better query more easily.

Multifacted search is an important paradigm for extracting and mining applications that provides users to analyze, navigate through multidimensional data.

Facetted search [91] can also be applied on spoken web search problem to index the metadata associated with audio content that provides audio search solution for rural people. The query interface ensures that a user is able to narrow the search results fastly. The approach focuses on indexing system and not generating precision - recall results on a labeled set of data.

Kong et al., [96] incorporated the feedback of users on the query facets into document ranking for evaluating boolean filtering feedback models that are widely used in conventional faceted search which automatically generates the facets for a user given query instead of generating for a complete corpus. The boolean filtering model is less effective than soft ranking models.

Bron et al., [97] proposed a novel framework by adding type filtering based on category information available in wikipedia. Combining a language modelling approach with heuristic based on wikipedia's external links, framework achieves high recall scores by finding homepages of top ranked entities. The model returns entities that have not been judged.

Navarro et al., [98] develops an automatic facet generation framework for an efficient document retrieval. To extract the facets a new approach is developed which is both domain independent and unsupervised. The approach generates multifaceted topic effectively. The subtopics in the text collection is not investigated.

Liu et al., [99] presented the study of exploring topical lead lag across corpora. Selecting which text corpus leads and which lags in a topic is a big challenge. Text pioneer, a visual analytic tool is introduced. The tool investigates lead lag across corpora from global to local level. Multiple perspectives of results are conveyed by two visualizations like global lead lag as hybrid tree, local lead lag as twisted ladder. Text pioneer donot analyze topics within each corpus and across corpora.

Jiang et al., [100] presented Cross Lingual Query Log Topic Model (CL-QLTM) to investigate query logs to derive the latent topics of web search data. The model incorporates different languages by collecting co-occurence relations and cross lingual dictionaries from query log. CL-QLTM is effective and superior in discovering latent topics. The model is not applied on statistical machine translation.

Cafarella et al., [101] exploited the interesting knowledge from webpages which consists of higher relevance to user when compared to traditional approach. The system records co-occurences of schema elements and helps user in navigating, creating synonyms for schema matching use.

Wordnet Domains text document. The queries given by the user is free text queries. Mapping keywords to different attributes and their values of a given entity is a challenging task. Castanet is simple and effective that achieves higher quality results than other automated category creation algorithms. WordNet is not exhaustive and few other mechanism is needed to improve coverage for unknown terms.

Pound et al., [102] proposed a solution that exploits user facetted search behaviour and structured data to find facet utility. The approach captures values and conditional values that provides attributes and values according to user preferences. Experiment results show that the approach is scalable and also outperforms popular commercial systems.

Table 3: Performance of Models to Extract Facets

Sl.no.	Authors	Model	Advantages	Disadvantages
1.	Dou *et al.,* (2016) [88]	Unique Similarity Model	Duplicate List are eliminated	While labelling facets do not identify similar items
2.	Efstathiades *et al.,* (2016) [92]	k Relevant Nearest Neighbour	Relevant point of interest is extracted	Relevance score is not considered
3.	Zhang *et al.,* (2016) [93]	Inverted Linear quadtree	Reduces search space	-
4.	Pripuzie *et al.,* (2015) [94]	Space partition Probing	Top K objects are identified quickly	Fixed bounded region is not incorporated
5.	Hon *et al.,* (2014) [95]	Space Efficient Framework	Robust	Multiple patterns are not handled

Altingovde et al., [103] demonstrate static index pruning technique by incorporating query views like document and term centric. The technique improves the quality of top ranked result. When the web pages changes frequently the original index is not updated.

Koutris et al., [104] proposed a framework for pricing the data based on queries. The polynomial time algorithm is executed for a conjunctive queries of large class and the result shows that the data complexity instance based determincy is CO NP complete. The framework do not explore interaction between pricing and privacy.

Lucchese et al., [105] proposed two methodology for extracting user tasks when they search for relevant data from search engine. The method identifies user query logs and further aggregate same kind of users tasks based on supervised and unsupervised approaches. The method is effective in detecting similar latent needs from a query log. Users task by task search behaviour is not represented in the model.

Liu et al., [106] developed a tool that automatically differentiate structured data from search results. A feature type based approach is introduced which identifies a valid features and evaluates the quality of features using exact and heuristics computation methods. The method achieves local optimality avoids dependency on random initialization. Result differentiation (whether the selected features is interest to users are not) is not incorporated.

Liu et al., [107] proposed matrix representation to discover collection of documents based on user interest. The multidimensional visualization is presented to overcome the difficulty for users to compare across different facet values. The approach further enables visual ordering based on facet values to support cross facet comparisons of items and also support users in exploring tasks. The intradocument details are unavailable and visual scalability is not incorporated.

Efstathiades et al., [92] presents Link of Interest (LOI) to improve the quality of users queries. K Relevant Nearest Neighbor(K-RNN) queries is based on query processing method is proposed to analyse LOI information to retrieve relevant location based point of interest as shown in Table 3. The method captures the relevance aspect of data. Relevance score is not computed.

Hon et al., [95] developed space efficient frame works for top k string retrieval problem that considers two metrics for relevance features which includes frequency and proximity. The threshold based approach on these metrics are also been used. Compact space and sufficient space indexes are derived that results index space and query time with significant robustness. The framework is robust but do not index an the cache oblivious model and also the index takes twice the size of the text. Multiple patterns are not handled.

Zhang et al., [94] proposed (SPP) Space Partition and Probing to keep track of object position and relevance to the query and also to find the vector space. Quality is achieved by using MMR which is one of the important diversification algorithm. The method identifies the next top K object very quickly. SPP helps in reducing object axis and also increases the performance. Fixed bounded region is not considered. Zhang et al., [93] proposed inverted linear quadtree index structure to accomplish both spatial and keyword based techniques to effectively decreases the search space. Spatial keyword queries having two disputes: top k spatial keyword search(TOPK-SK) and batch top k spatial keyword search(BTOPK-SK), in which top-sk fetch the closest k objects which contains all keywords in the query. BTOPK-SK contains set of top k queries. Existing techniques in IL-quadtree presents firstly Keyword first index, which is to extract the related inverted indexes. Partition based method is proposed to further enhance the filtering capabilities of the signature of linear quadtree.

Catallo et al.,[108] proposed probabilistic k- Sky band to process subset of sliding window objects, that are most recent data objects. The algorithm out performs for parameter of large values of K parameter both in memory consumption and time reduction. Adaptive top K processing is not incorporated in the approach.

Bast et al.,[109] presented pre-processing techniques to achieve interactive query times on large text collections. Two similarities measures are considered which includes, firstly, query terms match -similar terms in collection. Secondly, Query terms match -terms with similar prefix in collection which display the results quickly and are more efficient and scalable.

Termehchy et al., [110] introduced the XML structure for searching the keyword effectively. Traditional keyword search techniques does not support effectively. In order to overcome these problems for data-centric, XML put forth the Coherency Ranking(CR), which is a database design self sustained ranking method for XML keyword queries that is based on prolonging concepts of data dependencies and mutual information. With the concepts of CR, that analyze the prior approaches to XML keyword search. Approximate coherency ranking and current potent algorithm process queries and rank their responds using CR. CR shows better precision and recall, provides better ranking than prior approaches.

Colini et al., [111] [112] proposes multiple keyword method that provides search auctions with budgets and bidders. Bidders is bounded by multiple slots per keyword. Bidders which have cumulative valuations are click through rates and budgets that confine the overall study of multiple keyword method. Multiple keywords mechanism is compatible, optimal and rational with expectation. In combinatorial setting, each bidder is having a direct involvement in a subset of keywords. Deterministic mechanisms with temper marginal valuations are incompatible.

Wu et al., [113] introduced the concept of safe zones. It studies the moving of top K keyword query. The safe zones saves the time and communication cost. The approach computes safe zone in order to optimize server side computation. It is also used to establish the client server communication. Spatial keyword is not processed and also the safe zone do not compute future path of moving the query.

Lu et al., [114] proposed reverse spatial keyword K nearest neighbour to find the query of object which is similar to one of the neighbour. The query search is based on spatial location and also text associated with it. The algorithm is used to prune unnecessary objects and also computes the lists. The method do not considers textual description of two different objects.

Cao et al., [115] demonstrates the concept of weighing a query. The spatial keywords match considers both the location and the text. The method focuses more on finding queries to group of objects by grouping spatial objects. Top K spatial keyword and weighing of query improves the performance and efficiency. The computational time is reduced but partial coverage of queries is not considered.

VI. Fine Grained Knowledge

Guan et al., [116] suggested "tcpdump" method to capture the web surfing activities from users. Web surfing activities reflects persons fine grained knowledge by recognizing the semantic structures. Further by using Dirichlet process infinite Guassian mixture model is adopted. D-iHMM process is employed for mining the fine grained aspect in each part by session clustering. Discovering fine grained knowledge reflected from people's interaction made knowledge sharing in collaborative environment much easier. Although privacy is major issue.

Wang et al., [117] analysed user's searching behaviors and considered inter-query dependencies. A semi-supervised clustering model is proposed based on the SVM framework. The model enables a more comprehensive understanding of user's searchbehaviors via query search logs and facilitates the development search-engine support for long-term tasks. The performance of the model is superior in identifying cross-session search. User modeling and long-term task based personalization is not considered.

Kotov et al., [118] proposed a method for creating a semi automatically labeled data set that can be used for identifying user's query searches from earlier sessions on the same task and to predict whether a user returns to the same task during his later session. Using logistic regression and MART classifiers the method can effectively model and analyze cross-session of user's information needs. The model is not incorporated in commercial search engines.

VII. Duplicate Detection and Data Fusion

Duplicate detection is the methodology of identification of multiple semantic representation of the existing and similar real world entities. The present day detection methods need to execute larger datasets in the least amount of time and hence to maintain the overall quality of datasets is tougher.

Papenbrock et al., [119] proposed a strategic approach namely the progressive duplicate detection methods as shown in Table 4 which finds the duplicates efficiently and reduces the overall processing time by reporting most of the results as shown in table 7 than the existing classical approaches.

Bano et al., [120] executed innovative windows algorithm that adapts window for duplicates and also which are not duplicates and unnecessary comparisons is avoided.

The duplicate records are a vital problem and a concern in knowledge management [124]. To Extract duplicate data items an entity resolution mechanism is employed for the procedure of cleanup. The overall evaluation reveals that the clustering algorithms perform extraordinarily well with accuracy and f-measure being high.

Whang et al., [125] investigates the enhancement of focusing on several matching records. Three types of hints that are compatible with different ER algorithms:(i) an ordered list of records, (ii) a sorted list of record pairs, (iii) a hierarchy of record partitions. The underlying disadvantage of the process is that it is useful only for database contents.

Duplicate records do not share a strategic key but they build duplicate matching making it a tedious task. Errors are induced because the results of transcription errors, incomplete information and lack of normal formats. Abraham et al., [126] [127] provides survey on different techniques used for detecting duplicates in both XML and relational data. It uses elimination rule to detect duplicates in database.

Elmagarmid et al., [128] present intensive analysis of the literature on duplicate record for detection and covers various similarity metrics, which will detect some duplicate records in exceedingly available information. The strengths of the survey analysis in statistics and machine learning aims to develop a lot of refined matching techniques that deem probabilistic models.

Deduplication is an important issue in the era of huge database [129]. Various indexing techniques have been developed to reduce the number of record pairs to be compared in the matching process. The total candidates generated by these techni- ques have high efficiency with scalability and have been evaluated using various data sets.

The training data in the form of true matches and true non matches is often unavailable in various real-world applications. It is commonly up to domain and linkage experts for decision of the blocking keys. Papadakis et al., [122] presented a blocking methods for clean-clean ER over Highly Heterogeneous Information Spaces (HHIS) through an innovative framework which comprises of two orthogonal layers. The effective layer incorporates methods for construction of several blockings with small probability of hits; the efficiency layer comprises of a rich variety of techniques which restricts the required number of pairwise matches.

Papadakis et al., [123] focuses to boost the overallblocking efficiency of the quadratic task on Entity Resolution among large, noisy, and heterogeneous information areas.

The problem of merging many large databases is often encountered in KDD. It is usually referred to as the Merge/Purge problem and is difficult to resolve in scale and accuracy. The Record linkage [130] is a well-known data integration strategy that uses sets for merging, matching and elimination of duplicate records in large and heterogeneous databases. The suffix grouping methodology facilitates the causal ordering used by the indexes

Table 4: Algorithms for Duplicate Detection

Sl.no.	Authors	Algorithm	Window selection	Advantages	Disadvantages
1.	Papenbrock *et al.*, (2015) [119]	PSNM	Adaptive	Efficient with limited execution time	Delivers results moderately
2.	Bano *et al.*, (2015) [120]	Innovative Window	Adaptive	Unnecessary Comparison is avoided	Do not support on Multiple Datasets
3.	Bronselaer *et al.*, (2015) [121]	Fusion Propogation	-	Conflicts in relationship attributes are resolved	More Expensive
4.	Papadakis *et al.*, (2013) [122]	Attribute Clustering	-	Effective on real world datasets	low quality blocks, Parallelizing is not adopted
5.	Papadakis *et al.*, (2011) [123]	-	Adaptive	Time Complexity is reduced	Process is very slow

for merging blocks with least marginal extra cost resulting in high accuracy. An efficient grouping similar suffixes is carried out with incorporation of a sliding window technique. The method is helpful in various health records for understanding patient's details but is not very efficient as it concentrates only on blocking and not on windowing technique. Additionally the methodology with duplicates that are detected using the state of the art expandable paradigm is approximate [131]. It is quite helpful in creating clusters records.

Bronselaer et al., [132] focused on Information aggregation approach which combine information and rules available from independent sources into summarization. Information aggregation is investigated in the context of inferencing objects from several entity relations. The complex objects are composed of merge functions for atomic and subatomic objects in a way that the composite function inherits the properties of the merge functions.

Sorted Neighborhood Method (SNM) proposed by Draisbach et al., [133] partitions data set and comparison are performed on the jurisdiction of each partition. Further, the advances in a window over the data is done by comparison of the records that appears within the range of same window. Duplicate Count Strategy (DCS) which is a variation of SNM is proposed by regulating the window size. DCS++ is proposed which is much better than the original SNM in terms of efficiency but the disadvantage is that the window size is fixed and is expensive for selection and operation. Some duplicates might be missed when large window are used.

The tuples in the relational structure of the database give an overview of the similar real world entities such tuples are described as duplicates. Deleting these duplicates and in turn facilitating their replacement with several other tuples represents the joint informational structure of the duplicate tuples up to a maximum level. The incorporated delete and then replacement mode of operation is termed as fusion. The removal of the original duplicate tuples can deviate from the referential integrity.

Bronselaer et al., [121] describes a technique to maintain the referential integrity. The fusion Propogation algorithm is based on first and second order fusion derivatives to resolve conflicts and clashes. Traditional referential integrity strategies like DELETE cascading, are highly sophisticated. Execution time and recursively calling the propagation algorithm increases when the length of chain linked relations increases.

Bleiholder et al., proposes the SQL Fuse by inducing the schema and semantics. The existential approach is towards the architecture, query languages, and query execution. The final step of actually aggregating data from multiple heterogeneous sources into a consistent and homogeneous datasetand is often inconsiderable.

Naumann et al., [134] observes that amount of noisy data are in abundance from several data sources. Without any suitable techniques for integrating and fusing noisy data with deviations, the quality of data associated with an integrated system remains extremely low. It is necessary for allowing tentative and declarative integration of noisy and scattered data by incorporating schema matching, duplicate detection and fusion. Subjected to SQL-like query against a series of tables instance, oriented schema matching covers the cognitive bridge of the varied tables by alignment of various corresponding attributes. Further, a duplicate detection technique is used for multiple representations of several matching entities. Finally, the paradigm of data fusion for resolving a conflict in turn merges around with each individualistic duplicate transforming it into a unique singular representation.

Bleiholder et al., [135] explains a conceptual understanding of classification of different operators over data fusion. Numerous techniques are based on standard and advanced operators of algebraic relations and SQL. The concept of Co-clustering is explained from several techniques for tapping the rich and associated meta tag information of various multimedia web documents that includes annotations, descriptions and associations. Varied Coclustering mechanisms are proposed for linked data that are obtained from multiple sources which do not matter the representational problem of precise texts but rather increase their performance up to the most minimally empirical measurement of the multi-modal features.

The two channel Heterogeneous Fusion ART (HF-ART) yields several multiple channels divergently. The GHF-ART [136] is designed to effectively represent multimedia content that incorporates Meta data to handle precise and noisy texts. It is not trained directly using the text features but can be identified as a key tag by training it with the probabilistic distribution of the tag based occurrences. The approach also incorporates a highly and the most adaptive methodology for active and efficient fusion of multimodal.

VIII. Conclusions

The paper presents different techniques and framework to extract relevant features from huge amount of unstructured text documents. The paper also reviews a survey on various text classification, clustering, summerization methods.

To guarantee the quality of extracted relevant features in a collection of text documents is a great challenge. Many text mining techniques have been proposed till date. However how effectively the discovered features is interesting and useful to the user is an open issue.

Our future work is to efficiently utilize relevant documents from non relevant documents. Effective filtering model is required to automatically generate facets. The security and time to extract the useful features that is duplicate free and fine grained knowledge helps the user to reduce time in searching various web pages needs to be addressed.

Références

1. R. Agrawal and M. Batra, “A Detailed Study on Text Mining Techniques,” International Journal of Soft Computing and Engineering (IJSCE) ISSN, vol. 2, no. 6, pp. 2231–2307, 2013.
2. V. H. Bhat, P. G. Rao, R. Abhilash, P. D. Shenoy, K. R. Venugopal, and L. Patnaik, “A Data Mining Approach for Data Generation and Analysis for Digital Forensic Application,” International Journal of Engineering and Technology, vol. 2, no. 3, pp. 313–319, 2010.
3. Y. Zhang, M. Chen, and L. Liu, “A Review on Text Mining,” In Proceedings of 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), pp. 681– 685, 2015.
4. S. Shehata, F. Karray, and M. S. Kamel, “An Efficient Concept-based Mining Model for Enhancing Text Clustering,” IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 1360–1371, 2010.
5. V. H. Bhat, P. G. Rao, R. Abhilash, P. D. Shenoy, K. R. Venugopal, and L. Patnaik, “A Novel Data Generation Approach for Digital Forensic Application in Data Mining,” In Proceedings of Second International Conference on Machine Learning and Computing (ICMLC), pp. 86–90, 2010.
6. D. E. Brown, “Text Mining the Contributors to Rail Accidents,” IEEE Transactions on Intelligent Transportation Systems, vol. 27, no. 5, pp. 1–10, 2015.
7. K. R. Venugopal, K. Srinivasa, and L. M. Patnaik, “Soft Computing for Data Mining Applications,” Springer, 2009.
8. V. K. Verma, M. Ranjan, and P. Mishra, “Text Mining and Information Professionals: Role, Issues and Challenges,” In Proceedings of 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services (ETTLIS), pp. 133–137, 2015.
9. A. Akilan, “Text mining: Challenges and Future Directions,” In Proceedings of Second International Conference on Electronics and Communication Systems (ICECS), pp. 1679–1684, 2015.
10. D. Sanchez, M. J. Martin-Bautista, I. Blanco, and C. Torre, “Text Knowledge Mining:An Alternative to Text Data Mining,” In Proceedings of IEEE International Conference on Data Mining Workshops(ICDMW), pp. 664–672, 2008.
11. Y. Dai, T. Kakkonen, and E. Sutinen, “Minedec:A Decision- Support Model that Combines Text-mining Technologies with Two Competitive Intelligence Analysis Methods,” International Journal of Computer Information Systems and Industrial Management Applications, vol. 3, no. 10, pp. 165–173,2011.
12. Y. Hu and X. Wan, “Ppsgen: Learning-Based Presentation Slides Generation for Academic Papers,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 4, pp. 1085– 1097, 2015.
13. E. D’Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, “Real-Time Detection of Traffic from Twitter Stream Analysis,” IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 4, pp. 2269–2283, 2015.
14. R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang, “Tedas: A Twitter-based Event Detection and

Analysis System," In Proceedings of IEEE 28th International Conference on Data Engineering (ICDE), pp. 1273–1276, 2012.
15. R. Parikh and K. Karlapalem, "Et: Events from Tweets," In Proceedings of the 22nd International Conference on World Wide Web Companion, pp. 613–620, 2013.
16. V. H. Bhat, V. R. Malkani, P. D. Shenoy, K. R. Venugopal, and L. Patnaik, "Classification of Email using Beaks: Behavior and Keyword Stemming," In Proceedings of IEEE Region 10 Conference TENCON, pp. 1139–1143, 2011.
17. A. Schulz, P. Ristoski, and H. Paulheim, "I See a Car Crash: Real-Time Detection of Small Scale Incidents in Microblogs," The Semantic Web: ESWC Satellite Events, pp. 22–33, 2013.
18. A. Gonzalez, L. M. Bergasa, and J. J. Yebes, "Text Detection and Recognition on Traffic Panels from Street-Level Imagery using Visual Appearance," IEEE Transactions on Intelligent Transportation Systems, vol. 15, no. 1, pp. 228–238, 2014.
19. D. P. Muni, N. R. Pal, and J. Das, "Genetic Programming for Simultaneous Feature Selection and Classifier Design," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 36, no. 1, pp. 106–117, 2006.
20. M. H. Aghdam, N. Ghasem-Aghaee, and M. E. Basiri, "Text Feature Selection Using Ant Colony Optimization," Expert Systems with Applications, vol. 36, no. 3, pp. 6843–6853, 2009.
21. K. Srinivasa, A. Singh, A. Thomas, K. R. Venugopal, and L. Patnaik, "Generic Feature Extraction for Classification using Fuzzy C-means Clustering," In Proceedings of 3rd International Conference on Intelligent Sensing and Information Processing, pp. 33–38, 2005.
22. E. Gasca, J. S. S´anchez, and R. Alonso, "Eliminating Redundancy and Irrelevance using a New Mlp-based Feature Selection Method," Pattern Recognition, vol. 39, no. 2, pp. 313–315, 2006.
23. R. K. Sivagaminathan and S. Ramakrishnan, "A Hybrid Approach for Feature Subset Selection using Neural Networks and Ant Colony Optimization," Expert systems with Applications, vol. 33, no. 1, pp. 49–60, 2007.
24. D. Cai, C. Zhang, and X. He, "Unsupervised Feature Selection for Multi-cluster Data," In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 333–342, 2010.
25. Z. Zhao, X. He, D. Cai, L. Zhang, W. Ng, and Y. Zhuang, "Graph Regularized Feature Selection with Data Reconstruction," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 689–700, 2016.
26. L. Xu, C. Jiang, Y. Ren, and H.-H. Chen, "Microblog Dimensionality Reduction—A Deep Learning Approach," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1779–1789, 2016.
27. D. Wang, F. Nie, and H. Huang, "Feature Selection via Global Redundancy Minimization," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 10, pp. 2743–2755, 2015.
28. H. Ogura, H. Amano, and M. Kondo, "Feature Selection with a Measure of Deviations from Poisson in Text Categorization," Expert Systems with Applications, vol. 36, no. 3, pp. 6826 – 6832, 2009.
29. N. Azam and J. Yao, "Comparison of Term Frequency and Document Frequency based Feature Selection Metrics in Text Categorization," Expert Systems with Applications, vol. 39, no. 5, pp. 4760–4768, 2012.
30. S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 1, pp. 40–51, 2007.
31. Z. Zhao, L. Wang, H. Liu, and J. Ye, "On Similarity Preserving Feature Selection," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, pp. 619–632, 2013.
32. Z. Zhao, X. He, L. Zhang, W. Ng, and Y. Zhuang, "Graph Regularized Feature Selection with Data Reconstruction," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 3, pp. 289–700, 2016.
33. J. Tang and H. Liu, "Unsupervised Feature Selection for Linked Social Media Data," In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 904–912, 2012.
34. H. Mousavi, D. Kerr, M. Iseli, and C. Zaniolo, "Harvesting Domain Specific Ontologies from Text," In Proceedings of IEEE International Conference on Semantic Computing (ICSC), pp. 211–218, 2014.
35. Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "l2, 1- norm Regularized Discriminative Feature Selection for Unsupervised Learning," In Proceedings of the International Joint Conference on Artificial Intelligence(IJCAI), vol. 22, no. 1, pp. 1589–1594, 2011.
36. D. Cai, X. He, J. Han, and T. S. Huang, "Graph Regularized Nonnegative Matrix Factorization for Data Representation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1548–1560, 2011.
37. D. Sejal, K. Shailesh, V. Tejaswi, D. Anvekar, K. R. Venugopal, S. Iyengar, and L. Patnaik, "Qrgqr: Query Relevance Graph for Query Recommendation," In Proceedings of IEEE Region 10 Symposium (TENSYMP), pp. 78–81, 2015.
38. W. Fan, N. Bouguila, and D. Ziou, "Unsupervised Hybrid Feature Extraction Selection for High-

Dimensional Non-Gaussian Data Clustering with Variational Inference," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 7, pp. 1670–1685, 2013.
39. V. Gupta and G. S. Lehal, "A Survey of Text Summarization Extractive Techniques," Journal of Emerging Technologies in Web Intelligence, vol. 2, no. 3, pp. 258–268, 2010.
40. P. S. Negi, M. Rauthan, and H. Dhami, "Text Summarization for Information Retrieval using Pattern Recognition Techniques," International Journal of Computer Applications, vol. 21, no. 10, pp. 20–24, 2011.
41. F. Debole and F. Sebastiani, "An Analysis of the Relative Hardness of Reuters-21578 Subsets," Journal of the American Society for Information Science and Technology, vol. 56, no. 6, pp. 584–596, 2005.
42. F. Xie, X. Wu, and X. Hu, "Keyphrase Extraction based on Semantic Relatedness," In Proceedings of 9th IEEE International Conference on Cognitive Informatics (ICCI), pp. 308– 312, 2010.
43. Y. Li, A. Algarni, and N. Zhong, "Mining Positive andNegative Patterns for Relevance Feature Discovery," In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 753–762, 2010.
44. N. Zhong, Y. Li, and S.-T. Wu, "Effective Pattern Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 30–44, 2012.
45. Y.-C. Chen, W.-C. Peng, and S.-Y. Lee, "Mining Temporal Patterns in Time Interval-Based Data," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 12, pp. 3318– 3331, 2015.
46. A. Bartoli, A. De Lorenzo, E. Medvet, and F. Tarlao, "Inference of Regular Expressions for Text Extraction from Examples," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 5, pp. 1217–1230, 2016.
47. Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. A. Bijaksana, "Relevance Feature Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1656–1669, 2015.
48. Q. Song, J. Ni, and G. Wang, "A Fast Clustering-based Feature Subset Selection Algorithm for High-Dimensional Data," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, pp. 1–14, 2013.
49. T.-S. Nguyen, H. W. Lauw, and P. Tsaparas, "Review Selection Using Micro-Reviews," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 4, pp. 1098–1111, 2015.
50. A. I. Kadhim, Y. Cheah, N. H. Ahamed, L. A. Salman et al., "Feature Extraction for Co-occurrence-based Cosine Similarity Score of Text Documents," In Proceedings of IEEE Student Conference on Research and Development (SCOReD), pp. 1–4, 2014.
51. D. Fradkin and D. Madigan, "Experiments with Random Projections for Machine Learning," In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 517–522, 2003.
52. S. Joshi, D. Shenoy, P. rashmi, K. R. Venugopal, and L. Patnaik, "Classification of Alzheimer's Disease and Parkinson's Disease by using Machine Learning and Neural Network Methods," In Proceedings of Second International Conference on Machine Learning and Computing (ICMLC), pp. 218–222, 2010.
53. H. Ahonen, O. Heinonen, M. Klemettinen, and A. I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," In Proceedings of IEEE International Forum on Research and Technology Advances in Digital Libraries, pp. 2–11, 1998.
54. X. Yan, J. Han, and R. Afshar, "Clospan: Mining Closed Sequential Patterns in Large Datasets," In SDM, pp. 166–177, 2003.
55. A. Gomariz, M. Campos, R. Marin, and B. Goethals, "Clasp: An Efficient Algorithm for Mining Frequent Closed Sequences," Advances in Knowledge Discovery and Data Mining, pp. 50–61, 2013.
56. J. Pei, J. Han, R. Mao et al., "Closet: An Efficient Algorithm for Mining Frequent Closed Itemsets," ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, vol. 4, no. 2, pp. 21–30, 2000.
57. J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, and M.-C. Hsu, "Freespan: Frequent Pattern-Projected Sequential Pattern Mining," In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 355–359, 2000.
58. J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.-C. Hsu, "Prefixspan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth," ICCN, pp. 215– 224, 2001.
59. K. R. Venugopal and R. Buyya, "Mastering c++," Tata McGraw-Hill Education, 2013.
60. M. N. Garofalakis, R. Rastogi, and K. Shim, "Spirit: Sequential Pattern Mining with Regular Expression Constraints," VLDB, vol. 99, pp. 7–10, 1999.
61. N. Zhong, Y. Li, and S.-T. Wu, "Effective Pattern Discovery for Text Mining," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, pp. 30–44, 2012.
62. A. Inje and U. Patil, "Operational Pattern Revealing Technique in Text Mining," In Proceedings of IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS), pp. 1–5, 2014.
63. R. J. Bayardo Jr, "Efficiently Mining Long Patterns from Databases," ACM Sigmod Record, vol. 27, no. 2, pp. 85–93, 1998.

64. Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," ICML, vol. 97, pp. 412–420, 1997.
65. P. D. Shenoy, K. Srinivasa, and L. M. K R Venugopal, Patnaik, "Dynamic Association Rule Mining using Genetic Algorithms," Intelligent Data Analysis, vol. 9, no. 5, pp. 439– 453, 2005.
66. M. Seno and G. Karypis, "Slpminer: An Algorithm for Finding Frequent Sequential Patterns using Length-Decreasing Support Constraint," In Proceedings of IEEE International Conference on Data Mining, pp. 418–425, 2002.
67. N. R. Mabroukeh and C. I. Ezeife, "A Taxonomy of Sequential Pattern Mining Algorithms," ACM Computing Surveys (CSUR), vol. 43, no. 1, pp. 1–41, 2010.
68. M. J. Zaki, "Spade: An Efficient Algorithm for Mining Frequent Sequences," Machine learning, vol. 42, no. 1-2, pp. 31– 60, 2001.
69. J. Han, J. Pei, Y. Yin, and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach," Data Mining and Knowledge Discovery, vol. 8, no. 1, pp. 53–87, 2004.
70. V. P. Raju and G. S. Varma, "Mining Closed Sequential Patterns in Large Sequence Databases," International Journal of Database Management Systems, vol. 7, no. 1, pp. 29–40, 2015.
71. J. Zhang, X. Zhao, S. Zhang, S. Yin, X. Qin, and I. Senior Member, "Interrelation Analysis of Celestial Spectra Data using Constrained Frequent Pattern Trees," Knowledge-Based Systems, vol. 41, no. 4, pp. 77–88, 2013.
72. F. Li, B. C. Ooi, M. T. O¨ zsu, and S. Wu, "Distributed Data Management using Mapreduce," ACM Computing Surveys (CSUR), vol. 46, no. 3, pp. 31–42, 2014.
73. N. Tiwari, S. Sarkar, U. Bellur, and M. Indrawan, "Classification Framework of Mapreduce Scheduling Algorithms," ACM Computing Surveys (CSUR), vol. 47, no. 3, pp. 49–88, 2015.
74. Y. Xun, J. Zhang, and X. Qin, "Fidoop: Parallel Mining of Frequent Itemsets Using Mapreduce," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, no. 3, pp. 313–325, 2016.
75. S. Sakr, A. Liu, and A. G. Fayoumi, "The Family of Mapreduce and Large-Scale Data Processing Systems," ACM Computing Surveys (CSUR), vol. 46, no. 1, pp. 11–44, 2013.
76. L. Wang, L. Feng, J. Zhang, and P. Liao, "An Efficient Algorithm of Frequent Itemsets Mining based on Mapreduce," Journal of Information and Computational Science, vol. 11, no. 8, pp. 2809–2816, 2014.
77. T. Ramakrishnudu and R. Subramanyam, "Mining Interesting Infrequent Itemsets from Very Large Data based on Mapreduce Framework," International Journal of Intelligent Systems and Applications, vol. 7, no. 7, pp. 44–64, 2015.
78. E. Ozkural, B. Ucar, and C. Aykanat, "Parallel Frequent Item Set Mining with Selective Item Replication," IEEE Transactions on Parallel and Distributed Systems, vol. 22, no. 10, pp. 1632–1640, 2011.
79. Z. Xu and R. Akella, "Active Relevance Feedback for Difficult Queries," In Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 459–468, 2008.
80. S. Desai, V. Chandrasheker, V. Mathapati, K. R. V. Rajuk, S. S. Iyengar, and L. M. Patnaik, "User Feedback Session with Clicked and Unclicked Documents for Related Search Recommendation," IADIS-International Journal on Computer Science and Information Systems, vol. 11, no. 1, pp. 81–98, 2016.
81. G. Cao, J.-Y. Nie, J. Gao, and S. Robertson, "Selecting Good Expansion Terms for Pseudo-Relevance Feedback," In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 243–250, 2008.
82. J. H. Paik, D. Pal, and S. K. Parui, "Incremental Blind Feedback: An Effective Approach to Automatic Query Expansion," ACM Transactions on Asian Language Information Processing (TALIP), vol. 13, no. 3, pp. 13–35, 2014.
83. A. Algarni, Y. Li, and Y. Xu, "Selected New Training Documents to Update User profile," In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 799–808, 2010.
84. S. Niharika, V. S. Latha, and D. Lavanya, "A Survey on Text Categorization," International Journal of Computer Trends and Technology, vol. 3, no. 1, pp. 39–45, 2012.
85. F. Sebastiani, "Machine Learning in Automated Text Categorization," ACM Computing Surveys (CSUR), vol. 34, no. 1, pp. 1–47, 2002.
86. R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes et al., "A Survey on Text Mining in Social Networks," The Knowledge Engineering Review, vol. 30, no. 2, pp. 157–170, 2015.
87. H. N. Vu, T. A. Tran, I. S. Na, and S. H. Kim, "Automatic Extraction of Text Regions from Document Images by Multilevel Thresholding and k-means Clustering," In Proceedings of IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS), pp. 329–334, 2015.
88. Z. Dou, Z. Jiang, S. Hu, J.-R. Wen, and R. Song, "Automatically Mining Facets for Queries from Their Search Results," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 2, pp. 385–397, 2016.

89. A. Sejal, K. Shailesh, V. Tejaswi, D. Anvekar, K. R. Venugopal, S. Iyengar, and L. Patnaik, “Query Click and Text Similarity Graph for Query Suggestions,” International Workshop on Machine Learning and Data Mining in Pattern Recognition, pp. 328–341, 2015.

90. X. Shi and C. C. Yang, “Mining Related Queries from Web Search Engine Query Logs using an Improved Association Rule Mining Model,” Journal of the American Society for Information Science and Technology, vol. 58, no. 12, pp. 1871–1883, 2007.

91. M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava, “Faceted Search and Browsing of Audio Content on Spoken Web,” In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1029–1038, 2010.

92. C. Efstathiades, A. Efentakis, and D. Pfoser, “Efficient Processing of Relevant Nearest-Neighbor Queries,” ACM Transactions on Spatial Algorithms and Systems (TSAS), vol. 2, no. 3, pp. 9–37, 2016.

93. C. Zhang, Y. Zhang, W. Zhang, and X. Lin, “Inverted Linear Quadtree: Efficient Top k Spatial Keyword S219219earch,” IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 7, pp. 1706–1721, 2016.

94. K. Pripuˇzi´c, I. P. ˇ Zarko, and K. Aberer, “Time - and Space-Efficient Sliding Window Top-k Query Processi- ng,” ACM Transactions on Database Systems (TODS), vol. 40, no. 1, pp. 1–36, 2015.

95. W.-K. Hon, R. Shah, S. V. Thankachan, and J. S. Vitter, “Space-Efficient Frameworks for Top-k String Retrieval,” Journal of the ACM (JACM), vol. 61, no. 2, pp. 9–45, 2014.

96. W. Kong and J. Allan, “Extending Faceted Search to the General Web,” In Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, pp. 839–848, 2014.

97. M. Bron, K. Balog, and M. De Rijke, “Ranking Related Entities: Components and Analyses,” In Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1079–1088, 2010.

98. G. Navarro, “Spaces, Trees, and Colors: The Algorithmic Landscape of Document Retrieval on Sequences,” ACM Computing Surveys (CSUR), vol. 46, no. 4, pp. 52–99, 2014.

99. S. Liu, Y. Chen, H. Wei, J. Yang, K. Zhou, and S. M. Drucker, “Exploring Topical Lead-Lag Across Corpora,” IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 1, pp. 115–129, 2015.

100. D. Jiang, Y. Tong, and Y. Song, “Cross-Lingual Topic Discovery from Multilingual Search Engine Query Log,” ACM Transactions on Information Systems (TOIS), vol. 35, no. 2, pp. 9–37, 2016.

101. M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, “Webtables: Exploring the Power of Tables on the Web,” In Proceedings of the VLDB Data to find Endowment, vol. 1, no. 1, pp. 538–549, 2008.

102. J. Pound, S. Paparizos, and P. Tsaparas, “Facet Discovery for Structured Web Search: A Query-Log Mining Approach,” In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 169–180, 2011.

103. I. S. Altingovde, R. Ozcan, and O¨ . Ulusoy, “Static Index Pruning in Web Search Engines: Combining Term and Document Popularities with Query Views,” ACM Transactions on Information Systems (TOIS), vol. 30, no. 1, pp. 2–30, 2012.

104. P. Koutris, P. Upadhyaya, M. Balazinska, B. Howe, and D. Suciu, “Query-Based Data Pricing,” Journal of the ACM (JACM), vol. 62, no. 5, pp. 43–87, 2015.

105. C. Lucchese, S. Orlando, R. Perego, F. Silvestri, and G. Tolomei, “Discovering Tasks from Search Engine Query Logs,” ACM Transactions on Information Systems (TOIS), vol. 31, no. 3, pp. 14–58, 2013.

106. Z. Liu and Y. Chen, “Differentiating Search Results on Structured Data,” ACM Transactions on Database Systems (TODS), vol. 37, no. 1, pp. 4–34, 2012.

107. V. Thai, P.-Y. Rouille, and S. Handschuh, “Visual Abstraction and Ordering in Faceted Browsing of Text Collections,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 3, no. 2, pp. 21–45, 2012.

108. Catallo, E. Ciceri, P. Fraternali, D. Martinenghi, and M. Tagliasacchi, “Top-k Diversity Queries Over Bounded Regions,” ACM Transactions on Database Systems (TODS), vol. 38, no. 2, pp. 10–55, 2013.

109. H. Bast and M. Celikik, “Efficient Fuzzy Search in Large Text Collections,” ACM Transactions on Information Systems (TOIS), vol. 31, no. 2, pp. 10–69, 2013.

110. A. Termehchy and M. Winslett, “Using Structural Information in Xml Keyword Search Effectively,” ACM Transactions on Database Systems (TODS), vol. 36, no. 1, pp. 4–44, 2011.

111. R. Colini-Baldeschi, S. Leonardi, M. Henzinger, and M. Starnberger, “On Multiple Keyword Sponsored Search Auctions with Budgets,” ACM Transactions on Economics and Computation, vol. 4, no. 1, pp. 2–36, 2016.

112. Arguello and R. Capra, “The Effects of Aggregated Search Coherence on Search Behavior,” ACM Transactions on Information Systems (TOIS), vol. 35, no. 1, pp. 2–32, 2016.

113. D. Wu, M. L. Yiu, and C. S. Jensen, “Moving Spatial Keyword Queries: Formulation, Methods, and Analysis,” ACM Transactions on Database Systems (TODS), vol. 38, no. 1, pp. 7–55, 2013.

114. Y. Lu, J. Lu, G. Cong, W. Wu, and C. Shahabi, “Efficient Algorithms and Cost Models for Reverse Spatial-Keyword K Nearest Neighbor Search,” ACM

Transactions on Database Systems (TODS), vol. 39, no. 2, pp. 13–61, 2014.

115. X. Cao, G. Cong, T. Guo, C. S. Jensen, and B. C. Ooi, "Efficient Processing of Spatial Group Keyword Queries," ACM Transactions on Database Systems (TODS), vol. 40, no. 2, pp. 13–59, 2015.

116. Z. Guan, S. Yang, H. Sun, M. Srivatsa, and X. Yan, "Fine- Grained Knowledge Sharing in Collaborative Environments," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 8, pp. 2163–2174, 2015.

117. H. Wang, Y. Song, M.-W. Chang, X. He, R. W. White, and W. Chu, "Learning to Extract Cross-Session Search Tasks," In Proceedings of the 22nd International Conference on World Wide Web, pp. 1353–1364, 2013.

118. A. Kotov, P. N. Bennett, R. W. White, S. T. Dumais, and J. Teevan, "Modeling and Analysis of Cross-Session Search Tasks," In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 5–14, 2011.

119. T. Papenbrock, A. Heise, and F. Naumann, "Progressive Duplicate Detection," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1316–1329, 2015.

120. H. Bano and F. Azam, "Innovative Windows for Duplicate Detection," International Journal of Software Engineering and Its Applications, vol. 9, no. 1, pp. 95–104, 2015.

121. A. Bronselaer, D. Van Britsom, and G. De Tre, "Propagation of Data Fusion," IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1330–1342, 2015.

122. G. Papadakis, E. Ioannou, T. Palpanas, C. Nieder´ee, and W. Nejdl, "A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 12, pp. 2665–2682, 2013.

123. G. Papadakis and W. Nejdl, "Efficient Entity Resolution Methods for Heterogeneous Information Spaces," In Proceedings of IEEE 27th International Conference on Data Engineering Workshops (ICDEW), pp. 304–307, 2011.

124. O. Hassanzadeh, F. Chiang, H. C. Lee, and R. J. Miller, "Framework for Evaluating Clustering Algorithms in Duplicate Detection," In Proceedings of the VLDB Endowment, vol. 2, no. 1, pp. 1282–1293, 2009.

125. S. E. Whang, D. Marmaros, and H. Garcia-Molina, "Pay-asyou- go Entity Resolution," IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 5, pp. 1111–1124, 2013.

126. A. A. Abraham and S. D. Kanmani, "A Survey on Various Methods used for Detecting Duplicates in Xml Data," International Journal of Engineering Research and Technology, vol. 3, no. 1, pp. 1–10, 2014.

127. J. J. Tamilselvi and C. B. Gifta, "Handling Duplicate Data in Data Warehouse for Data Mining," International Journal of Computer Applications (0975–8887), vol. 15, no. 4, pp. 1–9, 2011.

128. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 1, pp. 1–16, 2007.

129. P. Christen, "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 9, pp. 1537– 1555, 2012.

130. T. De Vries, H. Ke, S. Chawla, and P. Christen, "Robust Record Linkage Blocking using Suffix Arrays and Bloom Filters," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 5, no. 2, pp. 9–44, 2011.

131. O. Hassanzadeh and R. J. Miller, "Creating Probabilistic Databases from Duplicated Data," The VLDB Journal—The International Journal on Very Large Data Bases, vol. 18, no. 5, pp. 1141–1166, 2009.

132. A. Bronselaer and G. De Tr´e, "Aspects of object Merging," Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), pp. 1–6, 2010.

133. U. Draisbach, F. Naumann, S. Szott, and O. Wonneberg, "Adaptive Windows for Duplicate Detection," In Proceedings of IEEE 28th International Conference on Data Engineering (ICDE), pp. 1073–1083, 2012.

134. F. Naumann, A. Bilke, J. Bleiholder, and M. Weis, "Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies." IEEE Data Engineering and Management, vol. 29, no. 2, pp. 21–31, 2006.

135. J. Bleiholder and F. Naumann, "Data Fusion," ACM Computing Surveys (CSUR), vol. 41, no. 1, pp. 1–41, 2009.

136. L. Meng, A.-H. Tan, and D. Xu, "Semi-Supervised Heterogeneous Fusion for Multimedia Data Co-Clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 9, pp. 2293–2306, 2014.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Review of Feature Selection and Optimization Strategies in Opinion Mining

By K. Venkata Rama Rao

Abstract- Opinion mining and sentiment analysis methods has become a prerogative models in terms of gaining insights from the huge volume of data that is being generated from vivid sources. There are vivid range of data that is being generated from varied sources. If such veracity and variety of data can be explored in terms of evaluating the opinion mining process, it could help the target groups in getting the public pulse which could support them in taking informed decisions. Though the process of opinion mining and sentiment analysis has been one of the hot topics focused upon by the researchers, the process has not been completely revolutionary. In this study the focus has been upon reviewing varied range of models and solutions that are proposed for sentiment analysis and opinion mining.

Keywords: opinion mining, sentiment analysis, social web data, machine learning, social media.

GJCST-C Classification: C.2.1,C.2.4, H.2.8

Review of Feature Selection and Optimization Strategies in Opinion Mining

K. Venkata Rama Rao

Abstract- Opinion mining and sentiment analysis methods has become a prerogative models in terms of gaining insights from the huge volume of data that is being generated from vivid sources. There are vivid range of data that is being generated from varied sources. If such veracity and variety of data can be explored in terms of evaluating the opinion mining process, it could help the target groups in getting the public pulse which could support them in taking informed decisions. Though the process of opinion mining and sentiment analysis has been one of the hot topics focused upon by the researchers, the process has not been completely revolutionary. In this study the focus has been upon reviewing varied range of models and solutions that are proposed for sentiment analysis and opinion mining. From the vivid range of inputs that are gathered and the detailed study that is carried out, it is evident that the current models are still in complex terms of evaluation and result fetching, due to constraints like comprehensive knowledge and natural language limitation factors. As a futuristic model in the domain, the process of adapting scope of evolutionary computational methods and adapting hybridization of such methods for feature extraction as an idea is tossed in this paper.

Keywords: opinion mining, sentiment analysis, social web data, machine learning, social media.

I. Introduction

Expressions of sentiments are analyzed using some of the prevalent methods like the sentiment analysis or the opinion mining machines. Profoundly the feelings that are generated in human mind is circumstantial on the basis of numerous thoughts, events, occurrence and reactions, which are mostly subjective in nature like mood, emotion. Predominantly the expressions and the state of mind of a human can be envisaged with the combination of certain factors like the emotions, mood and also the bodily gestures like the facial expressions and postures, alongside some of the communication forms.

Opinions and the expressions have significant importance in the day to day living, and focusing on such expressions and opinions have become an integral part of communication and living. With the advent development of social media and digital communication trends, voicing of opinions over such communication medium and people comprehending such inputs has become a norm. Also, with the emergence of BI solutions, and the analytics in place, the emphasis on what is being expressed in the social networking and other online medium by the people.

Pertaining to their likes and dislikes, views and choices has led to conditions, where the opinion mining is gaining significant importance.For instance, the companies are profoundly relying upon such insights for gathering customer views on the brand and the products of the organization, which can be resourceful to them in their strategic planning [1].

In economic factors evaluation, for interpretations that are apart from the technical knowledge and fundamentals based analysis, focusing on diverse range of inputs like the reforms, impending announcements, response to market trends, emerging market conditions, surge or deflation in prices, and many other such factors becomes significant for effective decisions.

In economics and finance to understand beyond fundamental and technical knowledge analysis, sentiment analysis supporters suggest additionally it is essential to use information as diverse as, impending announcements, sudden surge in commodity prices, rumors and reports of a market collapse or break through, increase in the interest rates by central banks, fluctuations in dollar prices, etc. as these factors help in better estimating and forecasting situations of changes in market. In Fig 1 the process flow of opinion mining and sentiment analysis is shown. Such inputs could be gathered from vivid range of sources. Sentiment analysis and opinion mining kind of solutions can be very resourceful in planning such development.

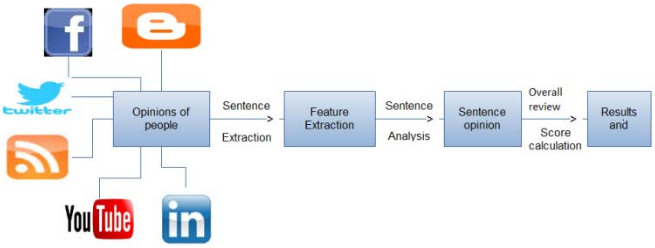


Figure 1: Sentiment Analysis or Opinion Mining model architecture

In the communication process, one of the critical aspects is about expressing about an action, a description in the combination of emotion, mood and sentiment. Also, an affect is also expressed with text and speech that are combined with the descriptions and using some language, which could be very resourceful

Author: B.Sc, MCA, B.Tech(CSE), M.Tech(CSE) Assistant Professor.
e-mail: kvrrao20@gmail.com

in sentimental analysis. Sam Glucksberg [2], in his book “understanding figurative language”, for instance, discuss a metaphor “my spouse’s lawyer is a shark and my job is my jail” in a characteristic manner, which express an effect in a suggestive and distinct manner.

In terms of expressing an effect, the descriptions [3] are usually used in daily language, which could be popular even in the domain of fictional content and is extend to scientific languages too [4], which are more effective in terms of natural [5]. Also the language in terms of biological sciences [6] is also at times adapted in the expression similar to the usage of some technical aspects pertaining finance and economics too.

Key purpose of a sentiment analysis system is to identify and extract the opinion of an author/speaker and find how astute the author is, towards giving value to something like a value, an object or event state or a concept. For handling the notions towards quantitative semantics, spatial relationships are adapted using the cognitive capabilities of a computing system for emulating a human mind. [7]. The key objective of performing such sentimental analysis is about detecting the instinctive emotional response of a reader for a speaker note and tabulate such responses for effective analysis.

II. State of The Art in Feature Selection and Optimization Strategies of Opinion Mining

In classification the removal of a relevant feature reduces the classification performance. The presence of irrelevant features also affects the classification performance [8]. Redundancy in between features is found by measuring the correlation between features [9] and removal of redundant features improves classification performance. In feature generation features numbering in the thousands are produced and generally a huge number of these features do not give any information because of class specific irrelevancy or redundancy and have to be removed to enhance the classification performance [10][11][8]. A feature is outstanding if it is highly discerning and improves the predictive capabilities of the classifier in assigning a new sample to a class [12] in sentiment analysis.

The performance of the machine learning models depends not only on efficient methods of feature extraction and weights assignment, but also on effective methods of feature selection. Feature selection objective is for selecting most relevant and discriminating features through the removal of features that are noisy and irrelevant for classification. If higher feature vector length and noisy and irrelevant features [13] are present the performance of the machine learning techniques is decreased. To reduce the dimensionality of the feature vector feature selection techniques such as information gain (IG) [14], mutual information (MI) [14], etc., or feature transformation techniques, like singular value decomposition (SVD), LDA, etc. are used. The techniques of feature selection select important features based on a term goodness value and selects top features if they are above a threshold criteria, and by dropping remaining irrelevant features. The techniques of feature transformation are based on converting feature vector of high-dimensionality into a feature space of low-dimensionality where the reduced feature vector similar to the earlier feature vector consists of the contributions of every feature. Wang and Wan [15] performed dimensionality reduction using the technique latent semantic analysis (LSA) and with the reduced feature vector applied the classifier support vector machine (SVM) to enhance the sentiment analysis performance. The popular approach for reducing feature vector length is feature selection compared to feature transformation due to its simplicity and computational efficiency.

A number of feature selection techniques such as MI, IG, DF (document frequency), CHI (chi-square), etc. [16][17][18][19] have been experimented for performing sentiment analysis. A most easy method of feature selection is DF (document frequency). Document Frequency is a regularly used sentiment analysis technique [20] based using the most frequently occurring terms in the texts that are used to construct a feature vector. Tan and Zhang [19] tested the performance of 4 techniques of feature selection (MI, IG, CHI, and DF) in the sentiment analysis of Chinese language documents. They used 5 machine learning algorithms, such as, centroid classifier, KNN (K-nearest neighbor), NB (naive Bayes), winnow classifier, and SVM (support vector machine). The experimental results show, the performance of IG in feature selection is better compared to the other approaches, and SVM algorithm performance is the best. Abbasi et al. [21] experimented with IG and GA (genetic algorithm) using a movie review dataset and proposed a hybrid approach called EWGA (entropy weighted genetic algorithm) to enhance sentiment analysis accuracy. Nicholls and Song [22] developed a new approach for feature selection called document frequency difference and performed comparison of the proposed approach with other techniques of feature selection for sentiment analysis. In [23] to perform sentiment analysis a log-likelihood approach is utilized for selecting key features.

Agarwal and Mittal [24] devised feature selection approach CPPD (Categorical Probability Proportion Difference) selects features based on their relevance and the features that are class discriminative. The feature selection method CPPD or Categorical Probability Proportion Difference combines two different techniques, Categorical Proportional Different (CPD) and Probability Proportion Difference (PPD). Categorical Proportional Different (CPD) technique is based on measuring the degree of contribution of a term in discriminating the class and in the classification only the topmost contributing terms are used [25]. The PPD technique is based on measuring the degree of belongingness/relatedness or probability that a term belongs to a specific class, and the difference is a measure of the ability to discriminate the class. In classification terms with higher degree of belongingness are chosen as candidate features. The advantage with CPD technique is it measures for a term its degree of class-distinguishing property that is a key attribute for a prominent feature. It is capable of removing terms that are unimportant for classification like terms that occur equally in both classes, and terms with high document frequency like stop words (i.e., a, the, an etc.). The advantage with PPD technique is that it is able to eliminate the terms with low document frequency, such as rare terms and are unimportant for sentiment analysis. Wang et al. [26] devised new Fisher's discriminant ratio-based feature selection technique for sentiment analysis of text data.

Duric and Song [27] introduced a novel feature selection technique based on a model of content and syntax that separates the reviewed entities and the opinion context (i.e., sentiment modifiers). The test performed with these features and using the maximum entropy classifier shows results comparable to that of state-of art methods. Agarwal and Mittal [17] approach for improving sentiment analysis performance is a hybrid approach of feature selection that combines the techniques, information gain and rough sets. Abbasi [28] devised feature selection approach intelligently explores and utilizes the syntactic and semantic information, and demonstrates how a heterogeneous feature set combined with a suitable technique of feature selection could enhance the sentiment analysis performance. O'keefe and Koprinska [29] proposed two techniques of feature selection, SentiWordNet Subjectivity Score (SWNSS) and SentiWordNet Proportional Difference (SWNPD). In sentiment analysis, SWNSS technique is capable of differentiating objective terms from subjective terms as only subjective terms may hold the sentiment, and SWNPD technique is capable of integrating feature selection with the ability to discriminate classes. Verma and Bhattacharyya [30] devised approach first prunes the data of terms that are semantically unimportant using a semantic score derived with SentiWordNet [31], and then information gain is used for selecting key features to enhance the accuracy of the sentiment classification.

Several feature selection techniques have been applied in the context of sentiment analysis such as, MI, DF, CHI, IG, etc. These techniques of feature selection have demonstrated to be efficient for performing sentiment analysis. They remove noise and redundancy occurring in the feature vector directly enhancing sentiment analysis performance in terms of execution time and accuracy. The main effort of the current models of feature selection is for selecting features relevant for the sentiment classification, and not the redundancy between the features.

a) *Heuristic Computation based Feature Selection Strategies*

Balahur et.al [32] has focused on comparative study of varied methods and resources that are used for mining opinions. Despite the level of complexities involved, the crux of the study is about the evaluation metrics adapted in terms of using the annotated quotations from varied news that are offered by EMS news gathering engine. The study concludes that the generic opinion mining systems needs the large lexicons and also the specialized training or test data which could make impact on the accuracy levels of the models.

Bo et.al [33] has reviewed model of techniques and approaches which shall directly support opinion-oriented information-seeking systems. The study focused on methods which seek new challenges that are raised by the sentiment-aware applications, when compared to the ones that already have traditional models of fact-based analysis.

In the process of review, some of the discussions related to available resources, datasets that are of benchmark and also the evaluation campaigns were also chosen. The availability and popularity of varied opinion-rich resources like the online review sites and also the personal blogs, emerging opportunities and the challenges that are creeping up from the extensive adaptation of ICT trends are utilized for understanding the opinions. The level of surge in the opinion mining and sentiment analysis manages with computation treatment of opinion, text subjectivity and sentiments.

Xu et.al [34] proposed model towards improving aspect-level opinion mining towards conducting online customer reviews. The new generative topic model of JAS (Joint Aspect/Sentiment) model was proposed to extract aspects and the aspect-dependent sentiment lexicons that could be derived from the online customer reviews. Among the aspect dependent sentiment lexicon indicates to aspect-specific opinion words comprising aspect-aware sentiment polarities pertaining to specific aspects. Also, the extracted aspect-dependent sentiment lexicons which are applied to opinion mining tasks at the aspect-level comprising various factors like aspect identification, aspect-based extractive opinion summarization and the aspect-level sentiment classification factors. Experimental study of the JAS model depicts the efficiency and effectiveness of the model in an intrinsic manner.

Emerging developments like the Web 2.0 and the social media content has triggered the scope of generating exhaustive range of content like the opinions, views, comments, reviews on varied socioeconomic factors, personal and lifestyle trends and market reviews. And the significant scope of focusing on such areas for sentiment analysis has been proposed by Zhang et.al [35].

If rightly adapted the online discussion forums, social networking postings and millions of tweets that are raised by the people can certainly support in gathering intrinsic insights in to varied emerging trends; Despite the fact that Opinion mining is a sub discipline of data mining models, still the level of computational and extraction techniques shall be very resourceful in gathering significant quantum of insightful inputs. Usage of sentiment analysis in the opinion mining can be very effective in terms of identifying the sentiment, affect and subjectivity, and also the emotional states using the online text. Zhai et.al [36] has proposed a model which focus on identifying product feature even before collecting the opinions of both positive and negative, for producing a summary of good or bad points. Aggregator sites of reviews and the ecommerce sites are some of the classic examples for business that is reliant on opinion mining for producing feature-based products quality summaries.

O'Keefe and Koprinska [29] have proposed the model of systematic evaluation for feature selectors and the feature weights using Naïve Bayes and the Support Vector Machine Classifiers. The mode inducts two new feature selection models and also proposes three feature weighting methods that could be adapted. Sentiment analysis focuses on whether the opinion in a document has to be positive or negative for a topic. The study also discuss that though there are many sentiment analysis applications that are proposed, still not many of them has the scope for handle large volume of features.

Abbasi et.al [21] has proposed the model of using the sentimental analysis methods for classifying the web for opinions in varied languages. For sentiment classification in both English and Arabic content, stylistic and syntactic feature extraction components are focused upon. Using the model of EWGA (Entropy Weighted Genetic Algorithm), a hybridized model of genetic algorithm for incorporating the information gain in terms of heuristic towards feature selection is also proposed in the study. It is imperative from the experimental study results of EWGA that it has outperformed in terms of features and techniques utility towards addressing document level sentiment classification.

Swaminathan et.al [37] proposed the model of using unstructured text being used as primary means for publishing biomedical research results. To extract and integrate data, text mining process can be applied in a routine manner. Some of the key aspects in the process is to extracting relationships between bio-entities like the foods and diseases. Also, the studies depict on stop short of how extracted relationships like the polarity and certainty levels could have impact.

Jeong et.al [38] has proposed the model of FEROM which focus on the process of feature extraction and refinement, and correct features towards reviewing data using exploitation of grammatical properties and also the semantic characteristic for feature words. The experimental studies of FEROM model depict that the process has been producing results in more accurate way and relative to functional opinion mining process.

Ozkis and Babalik [39] sounded the model of A-ABC (Accelerated Artificial Bee Colony) in which the two modifications are predominantly used on ABC algorithm for ensuring local search availability and also the levels of convergence speed. Modifications in the stream are depicted as Step Size (SS) Modification Rate (MR). Performances of A-ABC and the standards ABC are compared to varied benchmarking functions have resulted with better efficiency and outcome.

In extension to the models many more models similar to the A-ABC model has evolved. For instance, the models like MABC (Modified Artificial Bee Colony) [40], and another novel feature selection method proposed in [41] [42] also focus on optimal feature subset configurations and has achieved effective levels of classification accuracy when compared to the benchmarking models.

b) *Feature Weighting and Representation Schemes*

Feature weighting methods are of high importance for the approaches based on machine learning. These methods assign weights to features in terms of the importance levels of sentiment for enhanced classification of the sentiments data. There are different methods for assigning weights like, term frequency (TF), term-frequency-inverse document frequency (TF-IDF), binary weighting scheme etc. [18]. The binary weighting scheme assigns a feature value equal to 1 in the presence of a term else a value of 0 is given. The TFIDF method assigns weights to every term in terms of how rarely the terms occur in the remaining documents [14].

A comparison of these methods shows that for a maximum number of topic-based text classifications, TF-IDF weighing method performs efficiently, however for sentiment analysis, binary weighting method outperforms the other frequency-based methods [20]. An explanation for this may be because people write reviews usually using different sentiment words of expression. E.g. a person reviews a camera as, "This Nikon camera is great. Picture quality is clear and it looks nice." In this review 3 different sentiment words are used that are "great," "clear," and "nice." There are very less chances that same review will be written by him using only a single sentiment word "great", "this Nikon camera is great. Picture quality is great and its looks are also great." So the presence/absence of a term has a higher importance more than its frequency of occurrence in sentiment analysis.

Deng et al. [43] devised a method to weight a term based on its importance in the document and the term importance in expressing sentiment for sentiment analysis based on supervised approach. Martineau and Finin [44] devised a new method of finding feature weights called Delta TF-IDF that measures feature importance of a term based on its class distinguishing property and the terms that are unevenly distributed between positive and negative classes are assigned more weights with evenly distributed feature assigned a zero value. If a feature is more unevenly distributed among classes, the feature must be having more importance. In the sentiment analysis Delta TF-IDF outperformed the methods of term frequency and TF-IDF weighting.

Dai et al. [45] approach highlights the sentiment feature through an increase of their weights. Additionally with the bagging technique multiple classifiers are built on several feature spaces and a combination of all of these creates one aggregate classifier that is used to enhance the sentiment analysis performance. Paltoglou and Thelwallm [46] produced a complete study of several weighting techniques used in sentiment analysis. In the experiments performed by them, the classic $tf - idf$ variant methods showed to improve the sentiment analysis performance. Authors in [47] developed an improved mutual information based feature weighting method that assigns terms weights using sentiment scores. In the tests performed by them MI based weighting method is seen to be having higher effectiveness over the other methods. A few researchers have worked on including the words positional information in the text. One such research work by Raychev and Nakov [48] introduced a novel language-independent method of calculating term weights that uses the word position and its possibility of being subjective. Further multinomial naive Bayes is applied using the position information of the word in a feature set. This approach is tested with a movie dataset and the objective is to show that a particular word based on its position of occurrence in a document may have varied subjective power. O'Keefe and Koprinska [29] explored several feature weighting methods such as, feature frequency, feature presence (FP), and TFIDF, and based on the words grouped by their SentiWordNet (SWN) values, devised three new feature weighing methods, SWN Word Score Group (SWN-SG), SWN Word Polarity Groups (SWN-PG), and SWN Word Polarity Sums (SWN-PS).

The objective of the research for devising different weighting methods is of utilizing the information of feature polarity for assigning feature weights. The literature has numerous weighting methods produced such as, TF, binary weighting scheme, TF-IDF, etc. and an evaluation of these various weighing methods for sentiment analysis has shown binary weighting method to be the best.

III. Conclusion

In this paper, emphasis is on methodical approach of reviewing the varied models of sentimental analysis and opinion mining solutions that were earlier proposed in the studies. With the emerging trends of digital communication trends, adaptation of ICT trends and huge volume of opinion related data that is generated, the focus on sentiment analysis and opinion has been high in terms of research that is carried out. Review of the models of feature selection strategies and optimization techniques that are adapted in the machine learning based sentiment analysis depict that more models has been focused in such dimension. It is imperative from the review that due to some of the constraints like comprehensiveness and the knowledge of the problem, performing the tasks of sentiment analysis are much complex. Also, the issues like natural language processing limitations are also the other factors that impact the outcome of sentiment analysis. Despite the fact that considerable developments have taken place in the domain, still there is potential scope for improvement. One of the potential solutions that could be considered is about using the scope of evolutionary computational methods and adapting hybridization of such methods for feature extraction towards developing effective sentiment analysis model.

References Références Referencias

1. Candillier, Laurent, Frank Meyer, and M. Boullé. 2007. Comparing state-of-the-art collaborative filtering systems. International conference on Machine Learning and Data Mining MLDM 2007, Leipzig/ Germany. Lecture Notes in Computer Science, 2007, Volume 4571/2007:548– 562. Berlin: Springer.
2. Glucksberg, Sam. 2001. Understanding figurative language: From descriptions to idioms. Oxford: Oxford University Press.
3. Goatly, Andrew. 1997. The language of descriptions. London: Routledge.
4. Miller, A.L. 1984. Imagery in scientific thought. Boston, Basel, Stuttgart: Birkhaeuser.
5. Pullman, B. 1988. The atom in the history of human thought. Oxford: Oxford University Press.
6. Verschuuren, G.M. 1986. Investigating the life sciences: An introduction to the philosophy of science. Oxford: Pergamon Press.
7. Hobbs, Jerry R.1990. Literature and cognition. Lecture notes, number 21, Center for the Study of Language and Information, Stanford, California.
8. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5(1): 1205–1224.
9. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-depen-

dency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8): 1226–1238.

10. Agarwal B, Mittal N (2013) Optimal feature selection for sentiment analysis. In: Proceedings of the 14th international conference on intelligent text processing and computational linguistics (CICLing 2013), vol 7817, no 1, Samos, pp 13–24.
11. Forman G (2003) An extensive empirical study of feature selection metrics for text classification. J Mach Learn Res 3(1): 1289–1305.
12. Hoque N, Bhattacharyya DK, Kalita JK (2014) MIFS-ND: a mutual information-based feature selection method. Expert Syst Appl 41(14): 6371–6385.
13. Aphinyanaphongs Y, Fu LD, Li Z, Peskin ER, Efstathiadis E, Aliferis CF, Statnikov A (2014) A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. J Assoc Inf Sci Technol 65(10): 1964–1987.
14. Manning CD, Raghvan P, Schutze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge.
15. Wang L, Wan Y (2011) Sentiment classification of documents based on latent semantic analysis. In: Proceedings of the international conference on advanced research on computer education, simulation and modeling (CESM), Wuhan, pp 356–361
16. Agarwal B, Mittal N (2012) Text classification using machine learning methods-a survey. In: Proceedings of the 2nd international conference on soft computing for problem solving (SocPros-2012), vol 236, no 1, Jaipur, pp 701–710.
17. Agarwal B, Mittal N (2013) Sentiment classification using rough set based hybrid feature selection. In: Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis (WASSA'13), NAACL-HLT, Atlanta, pp 115–119.
18. Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and trends in information retrieval, vol 2, no 1–2. Now Publishers, Hanover, pp 1–135.
19. Tan S, Zhang J (2008) An empirical study of sentiment analysis for chinese documents. Expert Syst Appl 34(4): 2622–2629.
20. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Prague, pp 79–86.
21. Abbasi A, Chen H, Salem A (2008) Sentiment analysis in multiple languages: feature selection for opinion classification in Web forums. ACM Trans Inf Syst 26(3): 1–34.
22. Nicholls C, Song F (2010) Comparison of feature selection methods for sentiment analysis. In: Proceedings of the 23rd Canadian conference on advances in artificial intelligence. LNCS, vol 6085, no 1, Ottawa, pp 286–289.
23. Gamon M (2004) Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In: Proceedings of the 20th international conference on computational linguistics, Geneva, pp 841–848.
24. Agarwal B, Mittal N (2012) Categorical probability proportion difference (CPPD): a feature selection method for sentiment classification. In: Proceedings of the 2nd workshop on sentiment analysis where AI meets psychology, COLING 2012, Mumbai, pp 17–26.
25. Simeon M, Hilderman R (2008) Categorical proportional difference: a feature selection method for text categorization. In: Proceedings of the 7th Australasian data mining conference, Glenelg, pp 201–208.
26. Wang S, Li D, Song S, Wei Y, Li H (2009) A feature selection method based on Fisher's discriminant ratio for text sentiment classification. In: Proceedings of the international conference on web information systems and mining (WISM), Shanghai, pp 88–97.
27. Duric A, Song F (2011) Feature selection for sentiment analysis based on content and syntax models. In: Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, ACL-HLT, Portland, pp 96–103.
28. Abbasi A (2010) Intelligent feature selection for opinion classification. IEEE Intell Syst 25(4): 75–79.
29. O'keefe T, Koprinska I (2009) Feature selection and weighting methods in sentiment analysis. In: Proceedings of the 14th Australasian document computing symposium, Sydney, pp 67–74.
30. Verma S, Bhattacharyya P (2009) Incorporating semantic knowledge for sentiment analysis. In: Proceedings of the international conference on natural language processing (ICON), Hyderabad, pp 1–6.
31. Esuli A, Sebastiani F (2006) SentiWordNet: a publicly available lexical resource for opinion mining. In: Proceedings of 5th international conference on language resources and evaluation (LREC), Genoa, pp 417–422.
32. Balahur, A., Steinberger, R., Goot, E. V. D., Pouliquen, B., and Kabadjov, M. "Opinion mining on newspaper quotations", In Web Intelligence and Intelligent Agent Technologies, Vol. 3, pp. 523-526, 2009.
33. Bo, Pang., and Lillian, Lee. "Opinion Mining on Newspaper Quotations", Foundations and Trends in Information Retrieval, Vol.2, No.1-2, pp. 1–135, 2008.
34. Xu, X., Cheng, X., Tan, S., Liu, Y., and Shen, H. "Aspect-level opinion mining of online customer

reviews", Communications, China, Vol.10, No.3, pp.25-41, 2013.

35. Zhang, Y., Yu, X., Dang, Y., and Chen, H. "An Integrated Framework for Avatar Data Collection from the Virtual World: A Case Study in Second Life", IEEE Transaction on Intelligent Systems, Vol.25, No.6, pp.17-23, 2010.
36. Zhai, Z., Liu, B., Wang, J., Xu, H., and Jia, P. "Product Feature Grouping for Opinion Mining", IEEE Transaction on Intelligent Systems, Vol.27, No.4, pp.37-44, 2012.
37. Swaminathan, R., Sharma, A., and Yang, H. "Opinion mining for biomedical text data: Feature space design and feature selection", In Proceedings of the 9th International Workshop on Data Mining in Bioinformatics, 2010.
38. Jeong, H., Shin, D., and Choi, J. "Ferom: Feature extraction and refinement for opinion mining", ETRI Journal, Vol.33, No.5,pp.720- 730, 2011.
39. Ozkis, A., and Babalik, A. "Accelerated ABC (A-ABC) Algorithm for Continuous Optimization Problems", Lecture Notes on Software Engineering, Vol.1, No.3, pp.262- 266,2013.
40. Anandhakumar, R., Subramanian, S., and Ganesan, S. "Modified ABC Algorithm for Generator Maintenance Scheduling", International Journal of Computer Science and Electrical Engineering, Vol.3, No.6, pp.812- 819, 2011.
41. Palanisamy, S., and Kanmani, S. "Artificial Bee Colony Approach for Optimizing Feature Selection", International Journal of Computer Science, Vol.9, No.3, pp. 432-438, 2012.
42. Karaboga, D., and Ozturk, C. "Fuzzy clustering with artificial bee colony algorithm", Scientific research and Essays, Vol.5, No.14, pp.1899- 1902, 2010.
43. Deng ZH, Luo KH, Yu HL (2014) A study of supervised term weighting scheme for sentiment analysis. Expert Syst Appl 41(7):3506–3513.
44. Martineau J, Finin T (2009) Delta TFIDF: an improved feature space for sentiment analysis. In: Proceedings of the third AAAI international conference on weblogs and social media, pp 258–261.
45. Dai L, Chen H, Li X (2011) Improving sentiment classification using feature highlighting and feature bagging. In: Proceedings of the 11th IEEE international conference on data mining workshops (ICDMW), Vancouver, pp 61–66.
46. Paltoglou G, Thelwallm M (2010) A study of information retrieval weighting schemes for sentiment analysis. In: Proceedings of the 48th annual meeting of the Association for Computational Linguistics, Uppsala, pp 1386–1395.
47. Lin Y, Zhang J, Wang X, Zhou A (2012) An information theoretic approach to sentiment polarity classification. In: Proceedings of the 2nd joint WICOW/AIRWeb workshop on web quality, Lyon, pp 35–40.
48. Raychev V, Nakov P (2009) Language-independent sentiment analysis using subjectivity and positional information. In: Proceedings of the international conference recent advances on natural language processing (RANLP), Borovets, pp 360–364.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

The Encryption Algorithms GOST28147-89-IDEA8-4 and GOST28147-89-RFWKIDEA8-4

By Gulom Tuychiev

National University of Uzbekistan

Abstract- In the paper created a new encryption algorithms GOST28147–89–IDEA8–4 and GOST28147–89–RFWKIDEA8– 4 based on networks IDEA8–4 and RFWKIDEA8–4, with the use the round function of the encryption algorithm GOST 28147–89. The block length of created block encryption algorithm is 256 bits, the number of rounds is 8, 12 and 16.

Keywords: *feystel network, lai–massey scheme, round function, round keys, output transformation, multipli-cation, addition, s–box.*

GJCST-C Classification: *E.3*

The Encryption Algorithms GOST28147–89–IDEA8–4 and GOST28147–89–RFWKIDEA8–4

Gulom Tuychiev

Abstract- In the paper created a new encryption algorithms GOST28147–89–IDEA8–4 and GOST28147–89–RFWKIDEA8–4 based on networks IDEA8–4 and RFWKIDEA8–4, with the use the round function of the encryption algorithm GOST 28147–89. The block length of created block encryption algorithm is 256 bits, the number of rounds is 8, 12 and 16.

Keywords: feystel network, lai–massey scheme, round function, round keys, output transformation, multiplication, addition, s–box.

I. Introduction

The encryption algorithm GOST 28147–89 [4] is a standard encryption algorithm of the Russian Federation. It is based on a Feistel network. This encryption algorithm is suitable for hardware and software implementation, meets the necessary cryptographic requirements for resistance and, therefore, does not impose restrictions on the degree of secrecy of the information being protected. The algorithm implements the encryption of 64–bit blocks of data using the 256 bit key. In round functions used eight S–box of size 4x4 and operation of the cyclic shift by 11 bits. To date GOST 28147–89 is resistant to cryptographic attacks.

On the basis of encryption algorithm IDEA and Lai–Massey scheme developed the networks IDEA8–4 [6] and RFWKIDEA8–4 [7], consisting from four round function. In the networks IDEA8–4 and RFWKIDEA8–4, similarly as in the Feistel network, in encryption and decryption using the same algorithm. In the networks used four round function having one input and output blocks and as the round function can use any transformation.

As the round function networks IDEA4–2 [1], RFWKIDEA4–2 [5], PES4–2 [8], RFWKPES4–2 [8], PES8–4 [2], RFWKPES8–4 [10], IDEA16–2 [11], RFWKIDEA16–2 [12] encryption algorithm GOST 28147–89 created the encryption algorithm GOST28147–89–IDEA4–2 [13], GOST28147–89–RFWKIDEA4–2 [14], GOST28147–89–PES4–2 [15], GOST28147–89–RFWKPES4–2 [16], GOST28147–89–PES8–4, GOST28147–89–RFWKPES8–4 [17], GOST28147–89–IDEA16–2, GOST28147–89–RFWKIDEA16–2 [18].

In this paper, applying the round function of the encryption algorithm GOST 28147–89 as round functions of the networks IDEA8–4 and RFWKIDEA8–4, developed new encryption algorithms GOST28147–89–IDEA8–4 and GOST28147–89–RFWKI DEA8–4. In the encryption algorithms GOST28147–89–IDEA8–4 and GOST28147–89–RFWKIDEA8–4 block length is 256 bits, the key length is changed from 256 bits to 1024 bits in increments of 128 bits and a number of rounds equal to 8, 12, 16, allowing the user depending on the degree of secrecy of information and speed of encryption to choose the number of rounds and key length. Below is the structure of the proposed encryption algorithm.

II. The Encryption Algorithm GOST28147–89–IDEA8–4

The structure of the encryption algorithm GOST28147–89–IDEA8–4. In the encryption algorithm GOST28147–89–IDEA8–4 length of the subblocks X^0, X^1, …, X^7, length of the round keys $K_{12(i-1)}$, $K_{12(i-1)+1}$, …, $K_{12(i-1)+7}$, $i=\overline{1...n+1}$, $K_{12(i-1)+8}$, $K_{12(i-1)+9}$, $K_{12(i-1)+10}$, $K_{12(i-1)+11}$, $i=\overline{1...n}$ and K_{12n+8}, K_{12n+9}, …, K_{12n+23} are equal to 32–bits. In this encryption algorithm the round function GOST 28147–89 is applied four time and in each round function used eight S–boxes, i.e. the total number of S–boxes is 32. The structure of the encryption algorithm GOST28147–89–IDEA8–4 is shown in Figure 1 and the S–boxes shown in Table 1.

Author: Technical Sciences (Ph.D.), National University of Uzbekistan. He received Pd.D. degree in specialty mathematic from the National University of Uzbekistan. e- mail: blasterjon@gmail.com

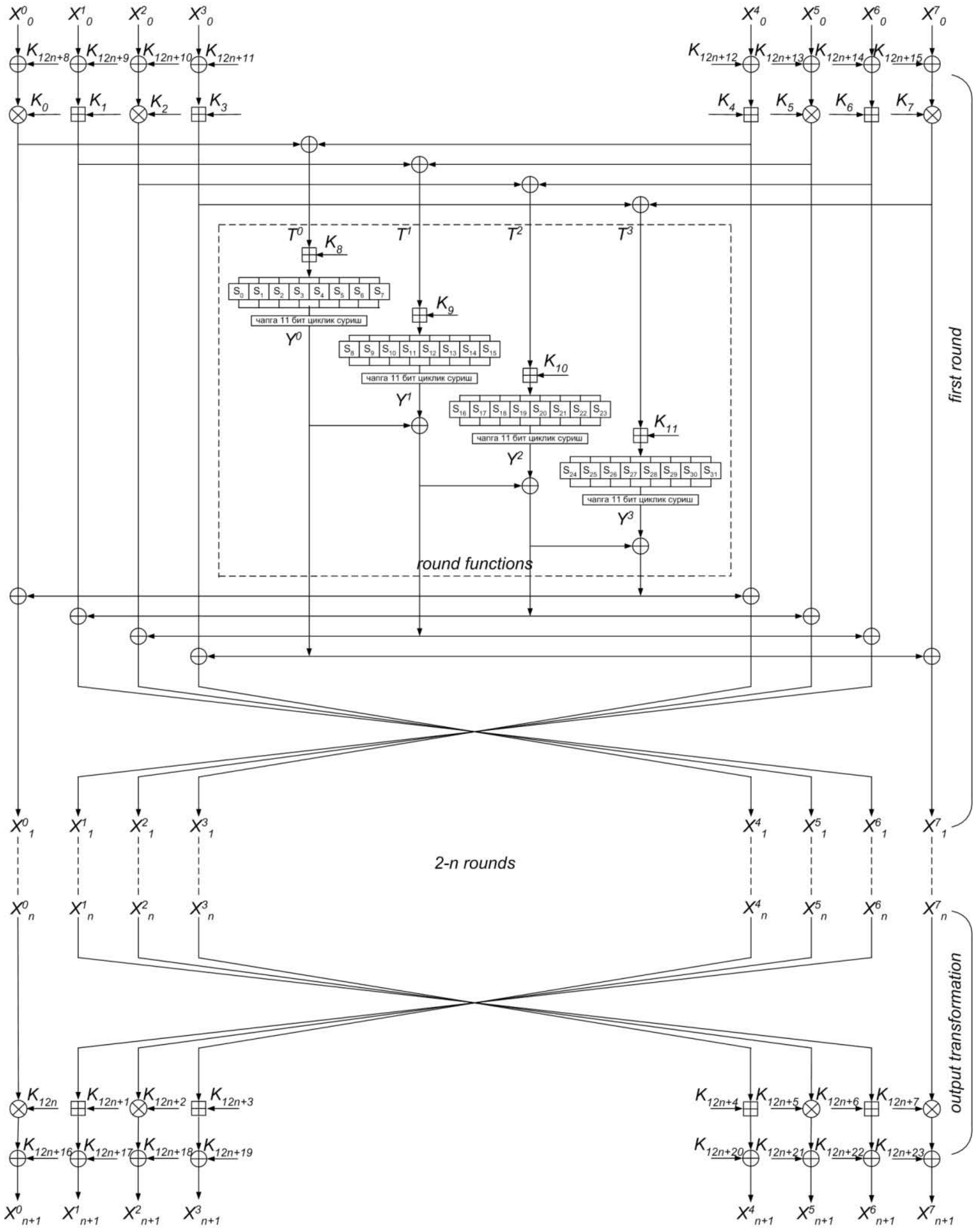


Figure 1: The scheme n–rounded encryption algorithm GOST28147–89–IDEA8–4

Consider the round function of a encryption algorithm GOST28147–89–IDEA8–4. The 32–bit subblocks T^0, T^1, T^2, T^3 are summed round keys $K_{12(i-1)+8}$, $K_{12(i-1)+9}$, $K_{12(i-1)+10}$, $K_{12(i-1)+11}$, $i=\overline{1...n}$, i.e. $S^0=T^0+K_{12(i-1)+8}$, $S^1=T^1+K_{12(i-1)+9}$, $S^2=T^2+K_{12(i-1)+10}$, $S^3=T^3+K_{12(i-1)+11}$. 32–bit subblocks S^0, S^1, S^2, S^3 divided into eight four–bit subblocks, i.e. $S^0=s_0^0\,||\,s_1^0\,||\,s_2^0\,||\,s_3^0\,||\,s_4^0\,||\,s_5^0\,||\,s_6^0\,||\,s_7^0$, $S^1=s_0^1\,||\,s_1^1\,||\,s_2^1\,||\,s_3^1\,||\,s_4^1\,||\,s_5^1\,||\,s_6^1\,||\,s_7^1$, $S^2=s_0^2\,||\,s_1^2\,||\,s_2^2\,||\,s_3^2\,||\,s_4^2\,||\,s_5^2\,||\,s_6^2\,||\,s_7^2$, $S^3=s_0^3\,||\,s_1^3\,||\,s_2^3\,||\,s_3^3\,||\,s_4^3\,||\,s_5^3\,||\,s_6^3\,||\,s_7^3$. The four–bit subblocks s_i^0, s_i^1, s_i^2, s_i^3, $i=\overline{0...7}$ transformed into the S–boxes: $R^0=S_0(s_0^0)\,||\,S_1(s_1^0)\,||\,S_2(s_2^0)\,||\,S_3(s_3^0)\,||\,S_4(s_4^0)\,||\,S_5(s_5^0)\,||\,S_6(s_6^0)\,||\,S_7(s_7^0)$, $R^1=S_8(s_0^1)\,||\,S_9(s_1^1)\,||\,S_{10}(s_2^1)\,||\,S_{11}(s_3^1)\,||\,S_{12}(s_4^1)\,||\,S_{13}(s_5^1)\,||\,S_{14}(s_6^1)\,||\,S_{15}(s_7^1)$, $R^2=S_{16}(s_0^2)\,||\,S_{17}(s_1^2)\,||\,S_{18}(s_2^2)\,||\,S_{19}(s_3^2)\,||\,S_{20}(s_4^2)\,||\,S_{21}(s_5^2)\,||$

$S_{22}(s_6^2)\,||\,S_{23}(s_7^2)$, $R^3 = S_{24}(s_0^3)\,||\,S_{25}(s_1^3)\,||\,S_{26}(s_2^3)\,||\,S_{27}(s_3^3)\,||\,S_{28}(s_4^3)\,||\,S_{29}(s_5^3)\,||\,S_{30}(s_6^3)\,||\,S_{31}(s_7^3)$. The resulting 32–bit subblocks R^0, R^1, R^2, R^3 cyclically shifted left by 11 bits and obtain subblocks Y^0, Y^1, Y^2, Y^3: $Y^0 = R^0 <<11$, $Y^1 = R^1 <<11$, $Y^2 = R^2 <<11$, $Y^3 = R^3 <<11$.

Table 1: The S–box of encryption algorithms

	0x0	0x1	0x2	0x3	0x4	0x5	0x6	0x7	0x8	0x9	0xA	0xB	0xC	0xD	0xE	0xF
S0	0x4	0x5	0xA	0x8	0xD	0x9	0xE	0x2	0x6	0xF	0xC	0x7	0x0	0x3	0x1	0xB
S1	0x5	0x4	0xB	0x9	0xC	0x8	0xF	0x3	0x7	0xE	0xD	0x6	0x1	0x2	0x0	0xA
S2	0x6	0x7	0x8	0xA	0xF	0xB	0xC	0x0	0x4	0xD	0xE	0x5	0x2	0x1	0x3	0x9
S3	0x7	0x6	0x9	0xB	0xE	0xA	0xD	0x1	0x5	0xC	0xF	0x4	0x3	0x0	0x2	0x8
S4	0x8	0x9	0x6	0x4	0x1	0x5	0x2	0xE	0xA	0x3	0x0	0xB	0xC	0xF	0xD	0x7
S5	0x9	0x8	0x7	0x5	0x0	0x4	0x3	0xF	0xB	0x2	0x1	0xA	0xD	0xE	0xC	0x6
S6	0xA	0xB	0x4	0x6	0x3	0x7	0x0	0xC	0x8	0x1	0x2	0x9	0xE	0xD	0xF	0x5
S7	0xB	0xA	0x5	0x7	0x2	0x6	0x1	0xD	0x9	0x0	0x3	0x8	0xF	0xC	0xE	0x4
S8	0xC	0xD	0x2	0x0	0x5	0x1	0x6	0xA	0xE	0x7	0x4	0xF	0x8	0xB	0x9	0x3
S9	0xE	0xF	0x0	0x2	0x7	0x3	0x4	0x8	0xC	0x5	0x6	0xD	0xA	0x9	0xB	0x1
S10	0xF	0xE	0x1	0x3	0x6	0x2	0x5	0x9	0xD	0x4	0x7	0xC	0xB	0x8	0xA	0x0
S11	0x1	0x8	0x7	0xD	0x0	0x4	0x3	0xF	0xB	0xA	0x9	0x2	0x5	0x6	0xC	0xE
S12	0x2	0xB	0x4	0xE	0x3	0x7	0x0	0xC	0x8	0x9	0xA	0x1	0x6	0x5	0xF	0xD
S13	0x3	0xA	0x5	0xF	0x2	0x6	0x1	0xD	0x9	0x8	0xB	0x0	0x7	0x4	0xE	0xC
S14	0x4	0x5	0xA	0x0	0xD	0x1	0x6	0x2	0xE	0x7	0xC	0xF	0x8	0x3	0x9	0xB
S15	0x5	0x4	0xB	0x1	0xC	0x0	0x7	0x3	0xF	0x6	0xD	0xE	0x9	0x2	0x8	0xA
S16	0x6	0x7	0x8	0x2	0xF	0x3	0x4	0x0	0xC	0x5	0xE	0xD	0xA	0x1	0xB	0x9
S17	0x7	0x6	0x9	0x3	0xE	0x2	0x5	0x1	0xD	0x4	0xF	0xC	0xB	0x0	0xA	0x8
S18	0x8	0x9	0x6	0xC	0x1	0xD	0xA	0xE	0x2	0xB	0x0	0x3	0x4	0xF	0x5	0x7
S19	0x9	0x8	0x7	0xD	0x0	0xC	0xB	0xF	0x3	0xA	0x1	0x2	0x5	0xE	0x4	0x6
S20	0xA	0xB	0x4	0xE	0x3	0xF	0x8	0xC	0x0	0x9	0x2	0x1	0x6	0xD	0x7	0x5
S21	0xB	0xA	0x5	0xF	0x2	0xE	0x9	0xD	0x1	0x8	0x3	0x0	0x7	0xC	0x6	0x4
S22	0xC	0xD	0x2	0x8	0x5	0x9	0xE	0xA	0x6	0xF	0x4	0x7	0x0	0xB	0x1	0x3
S23	0xD	0xC	0x3	0x9	0x4	0x8	0xF	0xB	0x7	0xE	0x5	0x6	0x1	0xA	0x0	0x2
S24	0x1	0x8	0x7	0x5	0x0	0xC	0xB	0xF	0x3	0x2	0x9	0xA	0xD	0x6	0x4	0xE
S25	0x2	0xB	0x4	0x6	0x3	0xF	0x8	0xC	0x0	0x1	0xA	0x9	0xE	0x5	0x7	0xD
S26	0x3	0xA	0x5	0x7	0x2	0xE	0x9	0xD	0x1	0x0	0xB	0x8	0xF	0x4	0x6	0xC
S27	0xF	0xE	0x1	0xB	0x6	0xA	0xD	0x9	0x5	0xC	0x7	0x4	0x3	0x8	0x2	0x0
S28	0xE	0xF	0x0	0xA	0x7	0xB	0xC	0x8	0x4	0xD	0x6	0x5	0x2	0x9	0x3	0x1
S29	0xA	0xB	0xC	0xE	0x3	0xF	0x0	0x4	0x8	0x1	0x2	0x9	0x6	0x5	0x7	0xD
S30	0xB	0xA	0xD	0xF	0x2	0xE	0x1	0x5	0x9	0x0	0x3	0x8	0x7	0x4	0x6	0xC
S31	0xC	0xD	0xA	0x8	0x5	0x9	0x6	0x2	0xE	0x7	0x4	0xF	0x0	0x3	0x1	0xB

Consider the encryption process of encryption algorithm GOST28147–89–IDEA8–4. Initially the 256–bit plaintext *X* partitioned into subblocks of 32–bits X_0^0, X_0^1, …, X_0^7 and runs the following steps:

1. subblocks X_0^0, X_0^1, …, X_0^7 summed by XOR with the keys K_{12n+8}, K_{12n+9}, …, K_{12n+15}: $X_0^j = X_0^j \oplus K_{12n+8+j}$, $j = \overline{0...7}$.
2. subblocks X_0^0, X_0^1, …, X_0^7 multiplied and summed with the round keys $K_{12(i-1)}$, $K_{12(i-1)+1}$, …, $K_{12(i-1)+7}$ and calculated 32–bit subblocks T^0, T^1, T^2, T^3 as follows: $T^0 = (X_{i-1}^0 \cdot K_{12(i-1)}) \oplus (X_{i-1}^4 + K_{12(i-1)+4})$, $T^1 = (X_{i-1}^1 \cdot K_{12(i-1)+1}) \oplus (X_{i-1}^5 + K_{12(i-1)+5})$, $T^2 = (X_{i-1}^2 \cdot K_{12(i-1)+2}) \oplus (X_{i-1}^6 + K_{12(i-1)+6})$, $T^3 = (X_{i-1}^3 \cdot K_{12(i-1)+3}) \oplus (X_{i-1}^7 + K_{12(i-1)+7})$, $i = 1$
3. to sublocks T^0, T^1, T^2, T^3 applying the round function and get the 32–bit subblocks Y^0, Y^1, Y^2, Y^3.
4. subblocks Y^0, Y^1, Y^2, Y^3 are summed to XOR with subblocks X_{i-1}^0, X_{i-1}^1, …, X_{i-1}^7, i.e. $X_{i-1}^0 = X_{i-1}^0 \oplus Y^3$, $X_{i-1}^1 = X_{i-1}^1 \oplus Y^2$, $X_{i-1}^2 = X_{i-1}^2 \oplus Y^1$, $X_{i-1}^3 = X_{i-1}^3 \oplus Y^0$, $X_{i-1}^4 = X_{i-1}^4 \oplus Y^3$, $X_{i-1}^5 = X_{i-1}^5 \oplus Y^2$, $X_{i-1}^6 = X_{i-1}^6 \oplus Y^1$, $X_{i-1}^7 = X_{i-1}^7 \oplus Y^0$, $i = 1$.
5. At the end of the round subblocks swapped, i.e, $X_i^0 = X_{i-1}^0$, $X_i^1 = X_{i-1}^6$, $X_i^2 = X_{i-1}^5$, $X_i^3 = X_{i-1}^4$, $X_i^4 = X_{i-1}^3$, $X_i^5 = X_{i-1}^2$, $X_i^6 = X_{i-1}^1$, $X_i^7 = X_{i-1}^7$, $i = 1$.

6. repeating the steps 2–5 *n* time, i.e. $i = \overline{2...n}$, obtained the subblocks X_n^0, X_n^1, …, X_n^7.
7. in output transformation round keys K_{12n}, K_{12n+1}, ..., K_{12n+7} are multiplied and summed into subblocks X_n^0, X_n^1, …, X_n^7, i.e. $X_{n+1}^0 = X_n^0 \cdot K_{12n}$, $X_{n+1}^1 = X_n^6 + K_{12n+1}$, $X_{n+1}^2 = X_n^5 \cdot K_{12n+2}$, $X_{n+1}^3 = X_n^4 + K_{12n+3}$, $X_{n+1}^4 = X_n^3 + K_{12n+4}$, $X_{n+1}^5 = X_n^2 \cdot K_{12n+5}$, $X_{n+1}^6 = X_n^1 + K_{12n+6}$, $X_{n+1}^7 = X_n^7 \cdot K_{12n+7}$.
8. subblocks X_{n+1}^0, X_{n+1}^1, ..., X_{n+1}^7 are summed by XOR with the round keys K_{12n+16}, K_{12n+17}, .., K_{12n+23}: $X_{n+1}^j = X_{n+1}^j \oplus K_{12n+16+j}$, $j = \overline{0...7}$.

As ciphertext receives the combined 32–bit subblocks $X_{n+1}^0 \| X_{n+1}^1 \| X_{n+1}^2 \| ... \| X_{n+1}^7$.

In the encryption algorithm GOST28147–89–IDEA8–4 when encryption and decryption using the same algorithm, only when decryption calculates the inverse of round keys depending on operations and are applied in reverse order. One important goal of encryption is key generation.

Key generation of the encryption algorithm GOST28147–89–IDEA8–4. In the n–round encryption algorithm GOST28147–89–IDEA8–4 used in each round 12 round keys of 32 bits and the output transformation of 8 round keys of 32 bits. In addition, prior to the first round and after the output transformation is applied 8 round keys on 32 bits. The total number of 32–bit round keys is equal to 12n+24. Hence, if n=8 then necessary 120, if n=12 then 168 and if n=16 then 216 to generate round keys.

The key of the encryption algorithm length of l ($256 \le l \le 1024$) bits is divided into 32–bit round keys K_0^c , K_1^c, ..., $K_{Lenght-1}^c$, $Lenght = l/32$, here $K = \{k_0, k_1, ..., k_{l-1}\}$, $K_0^c = \{k_0, k_1, ..., k_{31}\}$, $K_1^c = \{k_{32}, k_{33}, ..., k_{63}\}$, ..., $K_{Lenght-1}^c = \{k_{l-32}, k_{l-31}, ..., k_{l-1}\}$. Then calculated $K_L = K_0^c \oplus K_1^c \oplus ... \oplus K_{Lenght-1}^c$. If $K_L = 0$ then as K_L selected 0xC5C31537, i.e. $K_L = 0xC5C31537$. Round keys K_i^c, $i = \overline{Lenght...12n+23}$ calculated as follows: $K_i^c = SBox0(K_{i-Lenght}^c) \oplus SBox1(RotWord32(K_{i-Lenght+1}^c))$ $\oplus K_L$. After each generation of round keys value K_L cyclically shifted left by 1 bit. Here *RotWord32()*–cyclic shift 32 bit subblock to the left by 1 bit, *SBox()*–convert 32–bit subblock in S–box and $SBox0(A) = S_0(a_0) \| S_1(a_1) \| S_2(a_2) \| S_3(a_3) \| S_4(a_4) \| S_5(a_5) \| S_6(a_6) \| S_7(a_7)$, $SBox1(A) = S_8(a_0) \| S_9(a_1) \| S_{10}(a_2) \| S_{11}(a_3) \| S_{12}(a_4) \| S_{13}(a_5) \| S_{14}(a_6) \| S_{15}(a_7)$, $A = a_0 \| a_1 \| a_2 \| a_3 \| a_4 \| a_5 \| a_6 \| a_7$ and a_i – the four–bit sub–block.

Decryption round keys are computed on the basis of encryption round keys and decryption round keys output transformation associate with of encryption round keys as follows:

$$(K_{12n}^d, K_{12n+1}^d, K_{12n+2}^d, K_{12n+3}^d, K_{12n+4}^d, K_{12n+5}^d, K_{12n+6}^d, K_{12n+7}^d) =$$
$$((K_0^c)^{-1}, -K_1^c, (K_2^c)^{-1}, -K_3^c, -K_4^c, (K_5^c)^{-1}, -K_6^c, (K_7^c)^{-1}).$$

Decryption round keys of the second, third and *n*–round associates with the encryption round keys as follows:

$$(K_{12(i-1)}^d, K_{12(i-1)+1}^d, K_{12(i-1)+2}^d, K_{12(i-1)+3}^d, K_{12(i-1)+4}^d, K_{12(i-1)+5}^d,$$
$$K_{12(i-1)+6}^d, K_{12(i-1)+7}^d, K_{12(i-1)+8}^d, K_{12(i-1)+9}^d, K_{12(i-1)+10}^d, K_{12(i-1)+11}^d) =$$
$$((K_{12(n-i+1)}^c)^{-1}, -K_{6(n-i+1)+6}^c, (K_{12(n-i+1)+5}^c)^{-1}, -K_{12(n-i+1)+4}^c,$$
$$-K_{12(n-i+1)+3}^c, (K_{6(n-i+1)+2}^c)^{-1}, -K_{12(n-i+1)+1}^c, (K_{12(n-i+1)+7}^c)^{-1},$$
$$K_{12(n-i)+8}^c, K_{12(n-i)+9}^c, K_{12(n-i)+10}^c, K_{12(n-i)+11}^c), i = \overline{2...n}.$$

Decryption keys of the first round associated with the encryption keys as follows:

$$(K_0^d, K_1^d, K_2^d, K_3^d, K_4^d, K_5^d, K_6^d, K_7^d, K_8^d, K_9^d, K_{10}^d, K_{11}^d) =$$
$$((K_{12n}^c)^{-1}, -K_{12n+1}^c, (K_{12n+2}^c)^{-1}, -K_{12n+3}^c, -K_{12n+4}^c, (K_{12n+5}^c)^{-1},$$
$$-K_{12n+6}^c, (K_{12n+7}^c)^{-1}, K_{12(n-1)+8}^c, K_{12(n-1)+9}^c, K_{12(n-1)+10}^c, K_{12(n-1)+11}^c).$$

Decryption round keys applied to the first round and after the conversion of the output associated with encryption keys as follows: $K_{12n+8+j}^d = K_{12n+16+j}^c$, $K_{12n+16+j}^d = K_{12n+8+j}^c$, $j = \overline{0...7}$.

III. The Encryption Algorithm GOST28147–89–RFWKIDEA8–4.

The structure of the encryption algorithm GOST28147–89–RFWKIDEA8–4. In the encryption algorithm GOST28147–89–RFWKIDEA8–4 length of the subblocks X^0, X^1, …, X^7, length of the round keys $K_{8(i-1)}$, $K_{8(i-1)+1}$, …, $K_{8(i-1)+7}$, $i = \overline{1...n+1}$, K_{8n+8}, K_{8n+5}, ..., K_{8n+23} are equal to 32–bits. In this encryption algorithm the round function GOST 28147–89 is applied four time and in each round function used eight S–boxes, i.e. the total number of S–boxes is 32. The structure of the encryption algorithm GOST28147–89–IDEA8–4 is shown in Figure 2 and the S–boxes shown in Table 1.

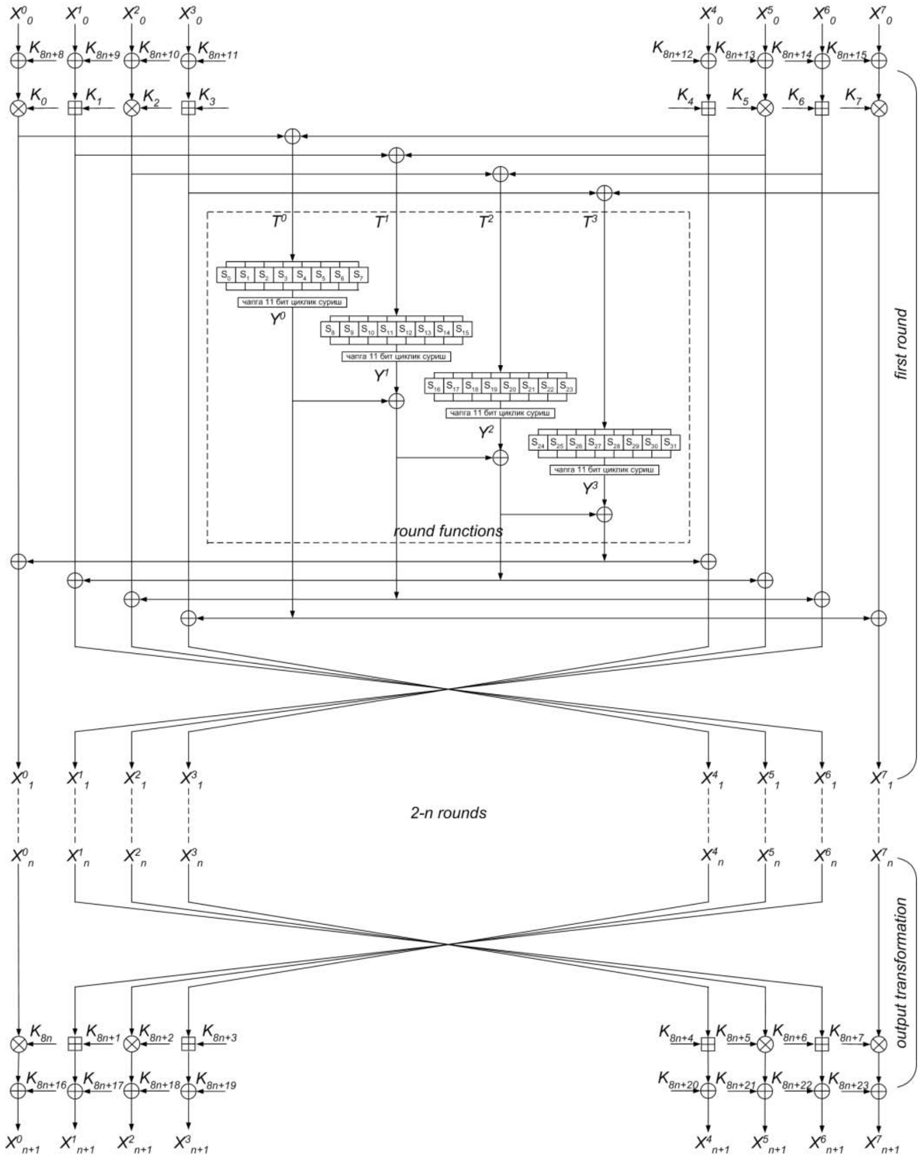


Figure 2: The scheme n–rounded encryption algorithm GOST28147–89–RFWKIDEA8–4

Consider the round function of encryption algorithm GOST28147–89–RFWKIDEA8–4. First 32–bit subblocks T^0, T^1, T^2, T^3 divided into eight four–bit sub–blocks, i.e. $T^0 = t_0^0 \| t_1^0 \| t_2^0 \| t_3^0 \| t_4^0 \| t_5^0 \| t_6^0 \| t_7^0$, $T^1 = t_0^1 \| t_1^1 \| t_2^1 \| t_3^1 \| t_4^1 \| t_5^1 \| t_6^1 \| t_7^1$, $T^2 = t_0^2 \| t_1^2 \| t_2^2 \| t_3^2 \| t_4^2 \| t_5^2 \| t_6^2 \| t_7^2$, $T^3 = t_0^3 \| t_1^3 \| t_2^3 \| t_3^3 \| t_4^3 \| t_5^3 \| t_6^3 \| t_7^3$. The four–bit subblocks t_i^0, t_i^1, t_i^2, t_i^3, $i = \overline{0...7}$ converted to S–box: $R^0 = S_0(t_0^0) \| S_1(t_1^0) \| S_2(t_2^0) \| S_3(t_3^0) \| S_4(t_4^0) \| S_5(t_5^0) \| S_6(t_6^0) \| S_7(t_7^0)$, $R^1 = S_8(t_0^1) \| S_9(t_1^1) \| S_{10}(t_2^1) \| S_{11}(t_3^1) \| S_{12}(t_4^1) \| S_{13}(t_5^1) \| S_{14}(t_6^1) \| S_{15}(t_7^1)$, $R^2 = S_{16}(t_0^2) \| S_{17}(t_1^2) \| S_{18}(t_2^2) \| S_{19}(t_3^2) \| S_{20}(t_4^2) \| S_{21}(t_5^2) \| S_{22}(t_6^2) \| S_{23}(t_7^2)$, $R^0 = S_{24}(t_0^3) \| S_{25}(t_1^3) \| S_{26}(t_2^3) \| S_{27}(t_3^3) \| S_{28}(t_4^3) \| S_{29}(t_5^3) \| S_{30}(t_6^3) \| S_{31}(t_7^3)$. Received 32–bit subblocks R^0, R^1, R^2, R^3 cyclically shifted to the left by 11 bits and get the subblocks Y^0, Y^1, Y^2, Y^3: $Y^0 = R^0 << 11$, $Y^1 = R^1 << 11$, $Y^2 = R^2 << 11$, $Y^3 = R^3 << 11$.

Consider the encryption process of encryption algorithm GOST28147–89–RFWKIDEA8–4. Initially the 256–bit plaintext X partitioned into subblocks of 32–bits X_0^0, X_0^1, ..., X_0^7 and performs the following steps:

1. sublocks X_0^0, X_0^1, ..., X_0^7 summed by XOR with the round keys K_{8n+8}, K_{8n+9}, ..., K_{8n+15}: $X_0^j = X_0^j \oplus K_{8n+8+j}$, $j = \overline{0...7}$.
2. sublocks X_0^0, X_0^1, ..., X_0^7 are multiplied and summed to the round keys $K_{8(i-1)}$, $K_{8(i-1)+1}$, ..., $K_{8(i-1)+7}$ and calculates a 32–bit subblocks T^0, T^1, T^2, T^3 as follows: $T^0 = (X_{i-1}^0 \cdot K_{8(i-1)}) \oplus (X_{i-1}^4 + K_{8(i-1)+4})$,

$$T^1 = (X^1_{i-1} \cdot K_{8(i-1)+1}) \oplus (X^5_{i-1} + K_{8(i-1)+5}),$$
$$T^2 = (X^2_{i-1} \cdot K_{8(i-1)+2}) \oplus (X^6_{i-1} + K_{8(i-1)+6}),$$
$$T^3 = (X^3_{i-1} \cdot K_{8(i-1)+3}) \oplus (X^7_{i-1} + K_{8(i-1)+7}),\ i = 1 .$$

3. to sublocks T^0, T^1, T^2, T^3 applying the round function and get the 32–bit subblocks Y^0, Y^1, Y^2 Y^3.
4. subblocks Y^0, Y^1, Y^2, Y^3 are summed to XOR with subblocks X^0_{i-1}, X^1_{i-1}, …, X^7_{i-1}, i.e. $X^0_{i-1} = X^0_{i-1} \oplus Y^3$ $X^1_{i-1} = X^1_{i-1} \oplus Y^2$, $X^2_{i-1} = X^2_{i-1} \oplus Y^1$, $X^3_{i-1} = X^3_{i-1} \oplus Y^0$ $X^4_{i-1} = X^4_{i-1} \oplus Y^3$, $X^5_{i-1} = X^5_{i-1} \oplus Y^2$, $X^6_{i-1} = X^6_{i-1} \oplus Y^1$ $X^7_{i-1} = X^7_{i-1} \oplus Y^0$, $i = 1$.
5. At the end of the round subblocks swapped, i.e, $X^0_i = X^0_{i-1}$, $X^1_i = X^6_{i-1}$, $X^2_i = X^5_{i-1}$, $X^3_i = X^4_{i-1}$ $X^4_i = X^3_{i-1}$, $X^5_i = X^2_{i-1}$, $X^6_i = X^1_{i-1}$, $X^7_i = X^7_{i-1}$, $i = 1$.
6. repeating the steps 2–5 *n* time, i.e. $i = \overline{2...n}$, obtained the subblocks X^0_n, X^1_n, …, X^7_n
7. in output transformation round keys K_{8n}, K_{8n+1}, ..., K_{8n+7} are multiplied and summed into subblocks X^0_n, X^1_n, …, X^7_n, i.e. $X^0_{n+1} = X^0_n \cdot K_{8n}$, $X^1_{n+1} = X^6_n + K_{8n+1}$, $X^2_{n+1} = X^5_n \cdot K_{8n+2}$, $X^3_{n+1} = X^4_n + K_{8n+3}$, $X^4_{n+1} = X^3_n + K_{8n+4}$, $X^5_{n+1} = X^2_n \cdot K_{8n+5}$, $X^6_{n+1} = X^1_n + K_{8n+6}$, $X^7_{n+1} = X^7_n \cdot K_{8n+7}$.
8. subblocks X^0_{n+1}, X^1_{n+1}, ..., X^7_{n+1} are summed by XOR with the round keys K_{8n+16}, K_{8n+17}, .., K_{8n+23}: $X^j_{n+1} = X^j_{n+1} \oplus K_{8n+16+j}$, $j = \overline{0...7}$.

As ciphertext receives the combined 32–bit subblocks $X^0_{n+1} \| X^1_{n+1} \| X^2_{n+1} \| ... \| X^7_{n+1}$.

In the encryption algorithm GOST28147–89–RFWKIDEA8–4 when encryption and decryption using the same algorithm, only when decryption calculates the inverse of round keys depending on operations and are applied in reverse order. One important goal of encryption is key generation.

Key generation of the encryption algorithm GOST28147–89–RFWKIDEA8–4. In the n–round encryption algorithm GOST28147–89–RFWKIDEA8–4 used in each round 8 round keys of 32 bits and the output transformation of 8 round keys of 32 bits. In addition, prior to the first round and after the output transformation is applied 8 round keys on 32 bits. The total number of 32–bit round keys is equal to 8n+24.

The key length of the encryption algorithm l ($256 \le l \le 1024$) bits is divided into 32–bit round keys K^c_0, K^c_1, …, $K^c_{Lenght-1}$, $Lenght = l/32$, here $K = \{k_0, k_1, ..., k_{l-1}\}$, $K^c_0 = \{k_0, k_1, ..., k_{31}\}$, $K^c_1 = \{k_{32}, k_{33}, ..., k_{63}\}$, ..., $K^c_{Lenght-1} = \{k_{l-32}, k_{l-31}, ..., k_{l-1}\}$. Then calculated $K_L = K^c_0 \oplus K^c_1 \oplus ... \oplus K^c_{Lenght-1}$. If $K_L = 0$ then as K_L selected 0xC5C31537, i.e. $K_L = 0\text{xC5C31537}$. Round keys K^c_i, $i = \overline{Lenght...8n+23}$ calculated as follows: $K^c_i = SBox0(K^c_{i-Lenght}) \oplus SBox1(RotWord32(K^c_{i-Lenght+1})) \oplus K_L$. After each generation of round keys value K_L cyclically shifted left by 1 bit.

Decryption round keys are computed on the basis of encryption round keys and decryption round keys of the first round associate with of encryption round keys as follows:

$$(K^d_0, K^d_1, K^d_2, K^d_3, K^d_4, K^d_5, K^d_6, K^d_7) = ((K^c_{8n})^{-1}, -K^c_{8n+1}, (K^c_{8n+2})^{-1}, -K^c_{8n+3}, -K^c_{8n+4}, (K^c_{8n+5})^{-1}, -K^c_{8n+6}, -K^c_{8n+7}).$$

Decryption round keys of the second, third and n–round associates with the encryption round keys as follows:

$$(K^d_{8(i-1)}, K^d_{8(i-1)+1}, K^d_{8(i-1)+2}, K^d_{8(i-1)+3}, K^d_{8(i-1)+4}, K^d_{8(i-1)+5}, K^d_{8(i-1)+6}, K^d_{8(i-1)+7}) = ((K^c_{8(n-i+1)})^{-1}, -K^c_{8(n-i+1)+6}, (K^c_{8(n-i+1)+5})^{-1}, -K^c_{8(n-i+1)+4}, -K^c_{8(n-i+1)+3}, (K^c_{8(n-i+1)+2})^{-1}, -K^c_{8(n-i+1)+1}, (K^c_{8(n-i+1)+7})^{-1}), i = \overline{2...n}.$$

Decryption keys output transformation associated with the encryption keys as follows:

$$(K^d_{8n}, K^d_{8n+1}, K^d_{8n+2}, K^d_{8n+3}, K^d_{8n+4}, K^d_{8n+5}, K^d_{8n+6}, K^d_{8n+7}) = ((K^c_0)^{-1}, -K^c_1, (K^c_2)^{-1}, -K^c_3, -K^c_4, (K^c_5)^{-1}, -K^c_6, (K^c_7)^{-1}).$$

Decryption round keys applied to the first round and after the conversion of the output associated with encryption keys as follows: $K^d_{8n+8+j} = K^c_{8n+16+j}$, $K^d_{8n+16+j} = K^c_{8n+8+j}$, $j = \overline{0...7}$.

IV. Results

As a result of this study built a new block encryption algorithms called GOST28147–89–IDEA8–4 and GOST28147–89–RFWKIDEA8–4. This algorithm is based on a networks IDEA16–2 and RFWKIDEA16–2 using the round function of GOST 28147–89. Length of block encryption algorithm is 256 bits, the number of rounds and key lengths is variable. Wherein the user depending on the degree of secrecy of the information and speed of encryption can select the number of rounds and key length.

It is known that S–box of the block encryption algorithm GOST 28147–89 are confidential and are used as long–term keys. In Table 2 below describes the options openly declared S–box such as: deg–degree of the algebraic nonlinearity; NL –nonlinearity; λ –relative resistance to the linear cryptanalysis; δ –relative resistance to differential cryptanalysis; SAC – criterion strict avalanche effect; the BIC criterion of independence of output bits. For S–box was resistant to crypt attack it is necessary that the values $\deg$ and NL were large, and the values λ, δ, SAC and BIC small.

Table 2: Parameters of the S–boxes of the GOST 28147–89

№	Parameters	S1	S2	S3	S4	S5	S6	S7	S8
1	$\deg$	2	3	3	2	3	3	2	2
2	NL	4	2	2	2	2	2	2	2
3	λ	0.5	3/4	3/4	3/4	3/4	3/4	3/4	3/4
4	δ	3/8	3/8	3/8	3/8	1/4	3/8	0.5	0.5
5	SAC	2	2	2	4	2	4	2	2
6	BIC	4	2	4	4	4	4	2	4

To S–Box was resistant to cryptanalysis it is necessary that the values $\deg$ and NL were large, and the values λ, δ, SAC and BIC small. In block cipher algorithms GOST28147–89–IDEA8–4 and GOST28147–89–RFWKIDEA8–4 for all S–boxes, the following equation: $\deg = 3$, $NL = 4$, $\lambda = 0.5$, $\delta = 3/8$, SAC=4, BIC=4. i.e. resistance is not lower than the algorithm GOST28147–89. These S–boxes are created based on Nyberg construction [3].

References Références Referencias

1. Aripov M., Tuychiev G.. The network IDEA4-2, consists from two round functions // Infocommuni cations: Networks–Technologies–Solutions. 2012, N4 (24), Tashkent", pp.55-59.
2. Aripov M., Tuychiev G.. The network PES8-4, consists from four round functions // Materials of the international scientific conference конференции «Modern problems of applied mathematics and information technologies-Al-Khorezmiy 2012», 2012, Vol.2, Tashkent, pp.16-19.
3. Bakhtiyorov U., Tuychiev G. About Generation Resistance S-Box And Boolean Function On The Basis Of Nyberg Construction // Materials scientific-technical conference «Applied mathematics and information security», Tashkent, 2014, 28–30 april, -pp. 317–324
4. GOST 28147-89. National Standard of the USSR. Information processing systems. Cryptographic protection. Algorithm cryptographic transformation.
5. Tuychiev G.. The networks RFWKIDEA4-2, IDEA4-1 and RFWKIDEA4-1 // Acta of Turin polytechnic university in Tashkent, 2013, N3, Tashkent, pp.71-77.
6. Tuychiev G.N. The network IDEA8–4, consists from four round functions // Infocommunications: Networks–Technologies–Solutions. –Tashkent, 2013, №2 (26), pp. 55–59.
7. Tuychiev G.N. About networks IDEA8–2, IDEA8–1 and RFWKIDEA8–4, RFWKIDEA8–2, RFWKIDEA8–1 developed on the basis of network IDEA8–4 // Uzbek mathematical journal, –Tashkent, 2014, №3, pp. 104–118
8. Tuychiev G.. The network PES4-2, consists from two round functions // Uzbek journal of the problems of informatics and energetics, 2013, N 5-6, Tashkent, pp.107-111.
9. Tuychiev G.. About networks PES4-1 and RFWKPES4-2, RFWKPES4-1 developed on the basis of network PES4-2 // Uzbek journal of the problems of informatics and energetics, 2015, N1-2, Tashkent, pp.100-105.
10. Tuychiev G.. About networks RFWKPES8-4, RFWKPES8-2, RFWKPES8-1, developed on the basis of network PES8-4 // Materials of the international scientific conference «Modern problems of applied mathematics and information technologies-Al-Khorezmiy 2014», 2014, vol.2, Samarkand, pp.32-36.
11. Tuychiev G.N. About networks IDEA16–4, IDEA16–2, IDEA16–1, created on the basis of network IDEA16–8 // Compilation of theses and reports republican seminar «Information security in the sphere communication and information. Problems and their solutions» –Tashkent, 2014
12. Tuychiev G.N. About networks RFWKIDEA16–8, RFWKIDEA16–4, RFWKIDEA16–2, RFWKIDEA16–1, created on the basis network IDEA16–8 // Ukrainian Scientific Journal of Information Security, –Kyev, 2014, vol. 20, issue 3, pp. 259–263
13. Tuychiev G.. Creating a data encryption algorithm based on network IDEA4-2, with the use the round function of the encryption algorithm GOST 28147-89 // Infocommunications: Networks-Technologies-Solutions, 2014, N4 (32), Tashkent, pp.49-54.
14. Tuychiev G.. Creating a encryption algorithm based on network RFWKIDEA4-2 with the use the round function of the GOST 28147-89 // International Conference on Emerging Trends in Technology, Science and Upcoming Research in Computer Science (ICDAVIM-2015), //printed in International Journal of Advanced Technology in Engineering and Science, 2015, vol.3, N1, pp.427-432.
15. Tuychiev G.. Creating a encryption algorithm based on network PES4-2 with the use the round function of the GOST 28147-89 // TUIT Bulleten", 2015, N4 (34), Tashkent, pp.132-136.

16. Tuychiev G.. Creating a encryption algorithm based on network RFWKPES4-2 with the use the round function of the GOST 28147-89 // International Journal of Multidisciplinary in Cryptology and Information Security, 2015, N2, vol.4, pp.14-17.
17. Tuychiev G.. The encryption algorithms GOST28147-89-PES8-4 and GOST28147-89-RFWKPES8-4 // «Information Security in the light of the Strategy Kazakhstan-2050»: proceedings III International scientific-practical conference (15-16 October 2015, Astana), 2015, Astana, pp.355-371.
18. Tuychiev G. The Encryption Algorithms GOST-IDEA16-2 and GOST-RFWKIDEA16-2 // Global journal of Computer science and technology: E Network, Web & security, vol 16, Issue 1, pp 30-38.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Evaluation of Features Extraction and Classification Techniques for Offline Handwritten Tifinagh Recognition

By Mouhcine Rabi, Mustapha Amrouch & Zouhir Mahani

Agadir, University Ibn Zohr

Abstract- This paper presents a review on different features extraction and classification methods for off-line handwritten Amazigh characters (called Tifinagh) recognition. The features extraction methods are discussed based on Statistical, Structural, Global transformation and moments.Although a number of techniques are available for feature extraction and classification,but the choice of an excellent technique decides the degree of accuracy of recognition. A series of experimentswere performed on AMHCD databaseallowing to evaluate the effectiveness of different techniques of extraction features based on Hidden Markov models, Neural network and Support vector Machine classifiers. The statistical techniques giveencouraging results.

Keywords : *handwritten recognition, tifinagh characters, extraction features (statistical, structural and global transformation), classification (HMM, MLP, SVM).*

GJCST-C Classification : *D.3.4,F.4.2*

Evaluation of Features Extraction and Classification Techniques for Offline Handwritten Tifinagh Recognition

Mouhcine Rabi α, Mustapha Amrouch σ & Zouhir Mahani ρ

Abstract- This paper presents a review on different features extraction and classification methods for off-line handwritten Amazigh characters (called Tifinagh) recognition. The features extraction methods are discussed based on Statistical, Structural, Global transformation and moments.Although a number of techniques are available for feature extraction and classification,but the choice of an excellent technique decides the degree of accuracy of recognition. A series of experimentswere performed on AMHCD databaseallowing to evaluate the effectiveness of different techniques of extraction features based on Hidden Markov models, Neural network and Support vector Machine classifiers. The statistical techniques giveencouraging results.

Keywords: handwritten recognition, tifinagh characters, extraction features (statistical, structural and global transformation), classification (HMM, MLP, SVM).

I. Introduction

Feature extraction in handwriting recognition is a very important field of image processing and object recognition. Fundamental component of characters are called features. The basic task of feature extraction and selection is to find out a group of the most effective features for classification; that is, compressing from high-dimensional feature space to low-dimensional feature space, so as to design classifier effectively.

Due to the nature of handwriting with its high degree of variability and imprecision obtaining these features, is a difficult task. Feature extraction methods are based on 3 types of features [1]:

- *Statistical:* Representation of a character image by statistical distribution of points takes care of style variations to some extent [2].
- *Structural:* Structural features are based on topological and geometrical properties of the character[3].
- *Global Transformations and Moments:* A continuous signal contains more information that can be represented for the purpose of classification [4].

In this paper our study was conducted to evaluate and examine the main approaches classes of extraction features on the Tifinagh script.

The majority of characters of this script are formed by loops, lines and curves (figure 1), this make it difficult to describe and sensitive to noise, the main problem is how to extracts features. This may be solved by the selection of the useful primitives customarily defined in the automatic character recognition.

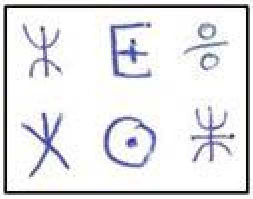


*Figure 1:*Some Tifinagh characters from AMHCD database [5]

Recently the recognition of handwritten Tifinagh characters (Figure 1) is the subject of several researches.These studies have been published in the literature. Among these researches, we find ([6][7][[8][9][10][11]).

All of previous cited works used a particular type of extraction features technique. [6]and [10] usedthe invariant moments aspattern sensitive features in classification and recognition. [7] and [8] used statistical techniques by applying respectively zoning method and freeman code to form the vector characteristics, whereas [9] used the Hough transform and the extracted features are structural based on the horizontal and vertical centreline of the letter in [11].

To evaluate the efficiency and the relevance of each type of extracted features we have used several methods of classification (Neural Networks, Hidden Markov Model and Support Vector Machine) for the recognition of Tifinagh characters.

The remainder of this paper is organized as follows. Section (2) presents the multiple techniques used to extract features from an image of Tifinagh letter after the preprocessing step. Section (3) is focused on the classification step. In section (4) we present the experimental results of several techniques used. The paper finally concludes with an analysis of the results and an introduction of future work.

Author α σ: Laboratory IRF-SIC, faculty of sciences, Ibn Zohr University Agadir, Morocco. e-mails: mouhcineh@gmail.com, m.amrouch@uiz.ac.ma
Author ρ: Hightschool of technology. IbnZohr University, Agadir, Morocco. e-mail: zouhir.mahani@uiz.ac.ma

II. Extraction Features

After a number of preprocessing operations such as binarization, noise reduction, skeletonization and normalization, a feature extraction method is applied to extract the most relevant characteristic of the character to recognize. The performance of a character recognition system largely depends on the quality and the relevance of the extracted features.

Features of a character can be classified into three main classes: Statistical features, Structural or topological features and Global transformations

a) Statistical Features

Statistical features are obtained from the arrangement of points constituting the character matrix. These features can be easily detected as compared to topological features. A number of techniques are used for feature extraction; some of these techniques used in this work are:

i. *Zoning*

Zoning According to this technique the character matrix is divided into small portions or zones (figure2 (a)). The densities of pixels in each zone are calculated and used as features; more details about zoning methods for handwritten character recognition are given in [14].

ii. *Diagonal based*

Diagonal features extraction[15][16] scheme for recognizing offline handwritten characters is proposed in this work. Every character image of size 100x100 is divided into 100 equal zones, each of size 10x10 pixels (figure 2(b))The features are extracted from each zone pixels by moving along the diagonals of its respective 10x10 pixels. Each zone has 19 diagonal lines and the foreground pixels present long each diagonal line is summed to get a single sub-feature, thus 19 sub-features are obtained from each zone, and then are averaged to form a single feature value placed in the corresponding zone(figure 4 (a)). This procedure is sequentially repeated for all zones. There could be some zones whose diagonals are empty of foreground pixels, the feature value corresponding to these zone are zero. Finally 100 features are extracted for each character figure 4(c).

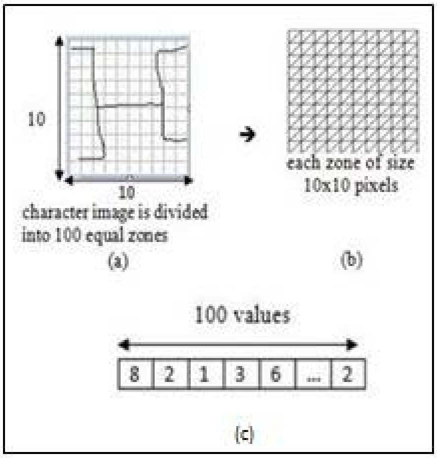

10
10
character image is divided into 100 equal zones
(a)
each zone of size 10x10 pixels
(b)
100 values
8 | 2 | 1 | 3 | 6 | ... | 2
(c)

Figure 2: Diagonal features based process

b) Structural or Topological features

Structural features are based on topological and geometrical properties of the character, such as aspect ratio, cross points, loops, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc.

In this study we used the Geometric features technique proposed in [17], this technique extracts the geometric features of the character contour. These features are based on the basic line types that form the character skeleton.

The image is divided into windows of equal size, and the feature is done on individual windows, for the system implemented, the image was zoned into equal sized windows.

To extract different line segments in a particular zone, the entire skeleton in that zone should be traversed. For this purpose, certain pixels in the character skeleton were defined as starters, intersections and minor starters (figure 3).

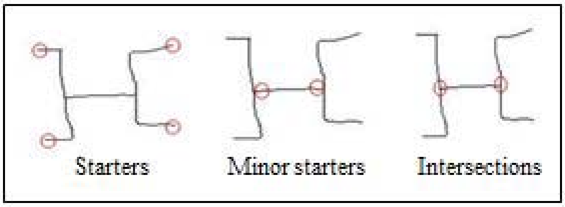


Figure 3: Starters, minor starters and intersections are rounded

After line segments have been extracted from the image, they have to be classified into any one of the following line types (Horizontal line, Vertical line, Right diagonal line, Left diagonal line).

After zonal feature extraction, certain features were extracted for the entire image based on the regional properties namely: Euler Number, Regional area, Eccentricity.

a) Moment and Global Transformations

The global transformations are generally widely used previously in the signal processing field. Their goals are change the image representation space(character or word)to facilitate the extraction of relevant features. There are many techniques used in handwritten recognition, in this work we have choice the Zernik moments and Gabor filter.

i. Zernike Moments

Zernike moments are used in pattern recognition applications as invariant descriptors of the image shape. They have been proven to be superior to moment functions such as geometric moments in terms of their feature representation capabilities and robustness in the presence of image quantization error and noise [18] They provide a compact way of describing an object's overall shape using a small set of values.

ii. Gabor Filter

Tifinagh character image features are extracted in this part using Gabor filters which can be written as a two dimensional Gabor function g(x, y), its Fourier transform G(u, v) as given in Equations below [20]:

$$g(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}\exp[-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})+2\pi jWx] \quad (1)$$

$$G(u,v)=\exp[-\frac{1}{2}(\frac{(u-W)^2}{\sigma_u^2}+\frac{v^2}{\sigma_v^2})] \quad (2)$$

Where σ_x, σ_y are the variances of x and y along the x, y axis, respectively; $\sigma u = ½ \pi \sigma x$ and $\sigma v = ½ \pi \sigma_y$.

After filtering the given input image, statistical features such as the mean and the variance of the image are computed. The extracted feature vector is constructed from the means and variances of all filtered images.

The Gabor filters are applied using the different orientations and scales. The mean μ and the standard deviation σ for each filtered image are then computed to form the character feature vector.

III. Classification

Classification is the process of assigning the sensed data to their corresponding class with respect to groups with homogeneous characteristics, with the aim of discriminating multiple objects from each other within the image. Some classification techniques used in this work are:

- *Neural Network (MLP):* The MLP is a special kind of Artificial Neural Network (ANN), is the mostly used classifier in the field of handwritten character recognition among the researcher [21].
- *Hidden Markov Model :*are a powerful tool frequently used in handwritten text recognition [22][23], and also in other fields related to pattern recognition and computational linguistics, like speech recognition, machine translation, Parts-Of-Speech tagging and information retrieval.
- *Support Vector Machines:*Support Vector Machines (SVMs) are a set of related supervised learning methods which can be used for both classification and regression [24].

IV. Experimental Results

A series of experiments have been performed to evaluate the effectiveness of different techniques of extraction features and classification. These experiments were performed on database of isolated Amazigh handwritten characters (AMHCD), 4200 character images from the portion of AMHCD were used in our experiment, 3100 character images were used for training and 930 character images were used to test identification performance.

The table below shows the experimental results of the different techniques of extraction features and classification's methods:

Table 1: Experimental results of extraction features techniques using various classifiers

		Extraction features techniques				
Classifier		Zoning	Diagonal	Geometric	Gabor	Zernike
NN	T.R	96.00	94.06	96.38	82.41	65.90
	R.R	82.04	86.75	74.62	71.39	42.15
HMM	T.R	75.87	81.55	76.26	57.20	51.03
	R.R	71.61	80.02	71.51	48.22	41.98
SVM	T.R	94.03	94.03	91.67	68.58	55.67
	R.R	85.59	**89.45**	78.17	68.06	47.09

T.R: Training Rate; R.R: Recognition Rate

The table shows the comparison of recognition rates between statistical, geometric, global transformations and Moments methods for extraction features using three divers classifier; NN,HMM and SVM.

As can be seen in table above, the results of recognition rate are varied according to extraction features technique used.

If we compared the results, we find that discrimination capability of statistical methods is better, whereas the Gabor filter and Zernike moments which are invariant to translation and rotation are limited for selection the pertinent features due to the similarity of Tifinagh characters (e.g.ⵇ and ⵄ, ⵟ and ⵃ).Structural technique gives an important resultsopening the way to a set of combination of statistic and geometric methods to integrate both the peculiarities of the text and the pixel distribution characteristics in the character image.

After analysing the result files that describe the target and actual outputs, we found that for some particular characters, the classification rate is poor. It can be explained that feature extraction techniques are influenced by many factors such as the variations of characters, the order of the strokes always different for different writers. Also, the form of the strokes can be varied. For example, the straight strokes can be curved as bows. Also the similarity of characters influence clearly the results, some characters were easily recognized as other particular characters such as the ⵣand ⵥ, ⵇ and ⵄ, ⵟ and ⵃ.

On other hand, the results are influenced mainly by the classifier performance, it is observed that recognition rate using HMMs are low compared to SVM and NN due the major problem of HMMs which is the estimation of emission probabilities, this confirms that HMMs are powerful to model sequences but still limited compared to NN and SVM in classification. To improve the results obtained using HMMs, it is recommended to use a hybrid classifier.

V. Conclusion

Feature extraction is an important phase in text recognition systems and for many pattern recognition problems.

In this paper, we have evaluated the feature extraction techniques for offline character recognition of Tifinagh script using various classifiers, the best recognition rate was achieved using statistical techniques. We noticed that the success rate of any recognition system depends not only on the features extraction but it depends on several reasons such as the recognizer technique, the pre-processing stage.

The work done is a first step for several perspectives. We try to improve the recognition rate by combining several classes of features to give a more general description of the character and classification techniques for a better representation and the speed of the system. We try to extend the approach to therecognition of words, sentences and texts and to other scripts, then exploit the results to develop a contextual recognition system.

References Références Referencias

1. B. El QacimyA. Hammouch ; M. A. Kerroum “A review of feature extraction techniques for hand written Arabic text recognition”. Electrical and Information Technologies (ICEIT), 2015 International Conference
2. S. Arora1 , D. Bhattacharjee2, M. Nasipuri2 , D. K. Basu2 , M.Kundu2 “Application of Statistical Features in Handwritten Devnagari Character Recognition” (2010)
3. S.A.Angadi and Sharanabasavaraj. H. Angadi “STRUCTURAL FEATURES FOR RECOGNITION OF HAND WRITTEN KANNADA CHARACTER BASED ON SVM” , International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol. 5,No.2, April 2015
4. J. H. AlKhateeb, R. Jinchang, J. Jianmin, S. S. Ipson and H. El-Abed, “Word-based Handwritten Arabic Scripts Recognition using DCT Features and Neural network Classifier”, In 5th International Multi-Conference on Systems, Signals and Devices, (2008), pp. 1–5.
5. Y. Es Saady, Ali Rachidi, Mostafa El Yassa and Driss Mammass, AMHCD: A Database for Amazigh Handwritten Character Recognition Research. International Journal of Computer Applications 27(4):44-, New York, USA August 2011.
6. Mohamed Abaynarh and Lahbib Zenkouar, “Offline Handwritten Characters Recognition Using Moments Features and Neural Networks”. Computer Technology and Application 6 (2015)
7. B. El Kessab, C. Daoui, B. Bouikhalene, R. Salouan“HandwritingMoroccanregionsrecognition using Tifinagh character” (2015)
8. A .HAIDAR, M.FAKIR, O.BENCHAREF Hybridation des modèles de Markov cachés et de la logique floue pour la reconnaissance des caractères Tifinagh manuscrits. 5ème conférence internationale sur les TIC pour l’amazighe 2012
9. Mustapha AMROUCH Reconnaissance des caractères imprimés et manuscrits, textes et documents basés sur les modèles de Markov cachés. Thèse de doctorat 2012

10. Rachid El Ayachi, Mohamed Fakir and Belaid Bouikhalene "Recognition of Tifinaghe Characters Using Dynamic Programming & Neural Net work" (2011)
11. Youssef Es Saady Amazigh Handwritten Character Recognition based on Horizontal and Vertical Centerline of Character (2011)
12. Vanita Mane, Leena Ragha, "Handwritten Character Recognition using Elastic Matching and PCA" International Conference on Advances in Computing, Communication and Control (ICAC 3'09) 2009 ACM ,410-415, 978-1-60558-351-8.
13. T.Y. Zhang and C.Y. Suen "A fast parallel Algorithm for Thinning Digital Patterns"Image processing and computer vision, 1984
14. D. Impedovo n, G.Pirlo "Zoning methods for handwritten character recognition: A survey" Pattern Recognition,Volume 47, Issue 3, March 2014, Pages 969–981, Handwriting Recognition and other PR Applications
15. A. Hirwan, S. Gonnade "Handwritten Character Recognition System Using Neural Network " International Journal of Advance Research in Computer Science and Management Studies .Volume 2, Issue 2, February 2014.
16. J.Pradeep,E.Srinivasan,S.Himavathi "Diagonal Feature Extraction Based Handwritten Character System Using Neural Network" International Journal of Computer Applications (0975 – 8887) Volume 8– No.9, October 2010
17. Dinesh Dileep A feature extraction technique based on character geometry for character recognition (2012)
18. Chee-Way Chong, P. Raveendran, R. Mukundan, A comparative analysis of algorithms for fast computation of Zernike moments. Pattern Recognition Journal volume 36, (2003) 731-742.
19. I. El-Fegh Handwritten Arabic Words Recognition using Multi Layer Perceptron and Zernik Moments
20. Hamdi Al-Jamimi and Sabri Mahmoud "Arabic Character Recognition Using Gabor Filters" Information and Computer Science, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.(2010)
21. Nibaran Das "Handwritten Arabic Numeral Recognition using a Multi Layer Perceptron" Computer Science and Engineering Department (2006)
22. Behrouz.Vaseghi1 and Somayeh.Hashemi Farsi Handwritten Word Recognition Using Discrete HMM and Self- Organizing Feature Map. 2012 International Congress on Informatics, Environment, Energy and Applications-IEEA 2012 IPCSIT vol.38 (2012) © (2012) IACSIT Press, Singapore
23. Grosicki E., El-Abed H., « ICDAR 2011: French Handwriting Recognition Competition », ICDAR, 2011.
24. César de Souza "Handwriting Recognition Revisited: Kernel Support Vector Machines", 2012

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

A Survey on Clustering Techniques for Multi- Valued Data Sets

By Lnc.Prakash K, K.Anuradha & D.Vasumathi

Annamacharya Institute of Technology and Sciences

Abstract- The complexity of the attributes in some particular domain is high when compare to the standard domain, the reason for this is its internal variation and the structure .their representation needs more complex data called multi-valued data which is introduced in this paper. Because of this reason it is needed to extend the data examination techniques (for example characterization, discrimination, association analysis, classification, clustering, outlier analysis, evaluation analysis) to multi-valued data so that we get more exact and consolidated multi-valued data sets. We say that multi-valued data analysis is an expansion of the standard data analysis techniques. The objects of multi-valued data sets are represented by multi-valued attributes and they contain more than one value for one entry in the data base. An example for this type of attribute is “languages known” .this attribute may contain more than one value for the corresponding objects because one person may be known more than one language.

GJCST-C Classification : H.3.3

A Survey on Clustering Techniques for Multi-Valued Data Sets

Lnc.Prakash K[α], K.Anuradha[σ] & D.Vasumathi[ρ]

Abstract- The complexity of the attributes in some particular domain is high when compare to the standard domain, the reason for this is its internal variation and the structure .their representation needs more complex data called multi-valued data which is introduced in this paper. Because of this reason it is needed to extend the data examination techniques (for example characterization, discrimination, association analysis, classification, clustering, outlier analysis, evaluation analysis) to multi-valued data so that we get more exact and consolidated multi-valued data sets. We say that multi-valued data analysis is an expansion of the standard data analysis techniques. The objects of multi-valued data sets are represented by multi-valued attributes and they contain more than one value for one entry in the data base. An example for this type of attribute is "languages known" .this attribute may contain more than one value for the corresponding objects because one person may be known more than one language.

I. Introduction

In the process of making more general surveys the persons are permitted to give more than one answer for a particular question. The answer may be a set of categorical values or a set of numerical values or sometimes it may be the combination of both. Now a days the data base is going on increasing and it is needed the integration of different data bases as a result a very lengthy databases are formed. The data bases that are formed in this way may contain the same objects that are repeated many times with different values .the increasing importance is to reduce the data base size by summarizing the data without loss of the information that is used for analysis. This can be achieved by introducing the concept of multi-valued attributes in the data bases. Because the cells of such data may contain not only single numerical or categorical values, but much more multifaceted information, such as subsets of categorical variable values, intervals of ordinal variable values, dependencies and need rules to be specified. These new statistical units are called multi-valued objects and there is a need for an extension of standard Data Analysis to such objects, called multi-valued Data analysis.

II. Input to Multi-Valued Data Analysis Algorithms

In general Columns of the source data table are variables (attributes) and rows (objects) are multi-valued descriptions. Each cell of this multi-valued data table may contain data of different types as given bellow.

- The sell may contain a quantitative value for example. If "height" is an attribute and a is a unit then height (a)=2.4 .
- The sell may contain discrete value, for instance, consider "village" is a variable and a is a unit then village (a) = Raja pet.
- The sell may contain Multi-valued, for example, in the quantitative case: height (a) =(3 .2, 4 .1,5)which means that the weight of a maybe 3 .2 or 4 .1 or 5 . In the discrete case, language (a) = (Telugu, Hindi, Tamil) means that the language of a may be Telugu or Hindi or Tamil. Here it is clear that the above two are special cases of this case.
- The sell may contain Interval valued data: for instance height (a) = [12, 15], which tells that the height of a varies in the interval [12, 15].

Some other types of variables may also exist with Multi valued weights, Taxonomic, Hierarchically dependent, logically dependent.

III. Result of Multi-Valued Data Analysis Algorithms

The output of any data analysis depending on the type of the data mining task. If the task is characterization it represents the nature of data and its character. When the task is association we can find the frequent items and then form the association rules. If the task is classification we can find the different classes that are associated with the data. If the task is clustering we can partition the data into different similar groups depending on the similarity measure.

IV. Source of Multi-Valued Data

Multi-valued data is formed from the tables that are integrated from different sources so this type of data is summarized and concise so the requirement of consolidation of different huge data sets causes multi-valued data sets. The result from several probability distributions, percentiles and the range of any random variables also produces multi-valued data. Some times

*Author α: Research scholar(JNTUH), Department of Computer Science and Engineering, AITS Rajampet, AP, India.
e-mail: klnc.prakash@gmail.com*
Author σ: Professor, Department of Computer Science and Engineering,GRIET, Hyderabad, TS, India. e-mail: kodali.anuradha@yahool.com
Author ρ: Professor, Department of Computer Science and Engineering, JNTUCEH, Hyderabad,TS, India. e-mail: rochan44@gmail.com

in order to answer to a query that is applied on different relational data bases form the multi-valued data sets. Multi-valued data can also be generated from Data Analysis (functional analysis, clustering, associations, neural networks . . .) of standard databases, from expert knowledge (scenario of traffic accidents, type of employed . . .), from time series (in the intervals of time), from private data (in the view of hiding the initial data by less precision), etc.

Table a

pan number	person's name	Age	category
11231	A.Ajay kumar	21.0	Male
14561	A.Vijay kumar	22.0	Male
17891	A.Arun kumar	23.0.	Male
11011	A.Varun kumar	24 .0	Male
11021	S.Swathi	25.0	Female
11031	S.Swetha	25.0	Female

Table b

pan number	City	Mode of payment	Date of payment	Amount in Rs
11231	Visakhapatnam(V)	Physical cash	1/06/13	500.00
11231	Hyderabad(H)	Credit to bank	10/07/13	80.00
11231	Nellore(N)	Check	11/07/13	100.00
14561	Visakhapatnam(V)	Physical cash	19/08/13	400.00
14561	Anakapalli(A)	Credit to bank	20/08/13	200.00
17891	Anakapalli(A)	Credit to bank	5/09/13	50.00
11011	Vizianagaram(Vz)	Physical cash	8/09/13	25.00

Table c

Name	Age	Category	Mode of payment	Amount in Rs	City
A.Ajay kumar	21.0	Male	Physical cash	500.00	Visakhapatanam(V)
A.Ajay kumar	21.0	Male	Credit to bank	80.00	Hyderabad(H)
A.Ajay kumar	21.0	Male	Check	100.00	Nellore(N)
A.Ajay kumar	22.0	Male	Physical cash	400.00	Visakhapatanam(V)
A.Ajay kumar	22.0	Male	Credit to bank	200.00	Anakapalli(A)
A.Arun kumar	23.0	Male	Credit	50.00	Anakapalli(A)
A.Varun kumar	24.0	Male	Physical cash	25.00	Vizianagaram(Vz)
A.Varun kumar	14.0	Male	Credit to bank	50.00	Visakhapatanam(V)

There are some approaches like ribeiro95 [1] and Thompson 91[5] identify the knowledge directly from domains that are structured. The most straight forward way to form a linear file is joining of related tables. the above table-c is the result of both of the tables table-a and table-b. Table-c is obtained by applying the natural join to both of the tables table-a and table-b which has the problem that a single object cannot be identified unique, that means a single object can be represented by more than one row. The problem with this type of representation is most of the data mining and machine learning algorithms identify each row as a unique object even though an object is identified by more than one tuple.

So, Therefore this construction between a structure database and linear file is understood by many analysis frame works in efficiently This paper introduces a data mining framework and a information discovery to deal with this drawback of the previous linear file representations and we finally concentrate on the issues of structured database and discovery of a set of queries that describe the features of objects in the database. To overcome from this problem mentioned above it is needed to develop a better illustration procedure which represents information for objects that are co-related with other objects.

One simple solution to solve the problem is to combine the associated objects into a single object by applying some aggregate operations. The issue in this concept is the selection of aggregate function that can be applied and minimizing the Loss of technical information because of the aggregate function so finally there forms a cluster of related objects that represent a set of objects as a single object. To solve the same problem another technique is introduction of multi-valued attributes for the objects. The generalization of this type of linear file arrangement is as multi-valued data set. in this types of data bases the values of multi-valued attribute are either a single value or a bin of values these bin of values represent the data that associate with the object with respect to the multi-valued attribute. Table-d consists the multi-valued data that is generated from the two tables table-a and table-b, in this table product type, purchase place and purchase amount are multi-valued attributes, the objects have a set of values in between braces in the corresponding cells of these multi-valued attributes, if there is no any value we represent it by null or if there is a single value we put that value as it is.

Table d: data set with multi-valued of attributes

Name	Age	Product type	category	Purchase place	Purchase amount
A.Ajay kumar	21.0	{physical cash,credit to bank,check}	Male	{Visakhapatnam(V),Hyderabad(H), Nellore(N)}	{500.00,80.00,100.00}
A.Vijay kumar	22.0	{ physical cash, credit to bank }	Male	{Visakhapatnam(V),anakapally(A)}	{400.00,200.00}
A.Arun kumar	23.0	{Credit credit to bank}	Male	Anakapally(A)	50.00
A.Varun kumar	24.0	{ physical cash, credit to bank }	Male	{vizianagaram(Vz),Visakhapatnam(V)}	{25.00,50.00}

Now table-d contains the unique tuples for representing the data of one person which is different from the data that is denoted by table-c. So a useful set of queries can be applied to this type of data for implementing data mining tasks, because of multi-valued attributes are existed in the data base the following problems are encountered during this process.

- It is needed to identify the set of queries that access the features of objects in the database.
- Applying the data mining techniques to the database that consists of multi-valued attributes.

Most of the data mining techniques that are existed cannot work properly on the multi-value datasets which consists of a set of values because they are developed for applying on single valued attributes. So the need is to develop the techniques that can be applied on multi valued attributes.

Multi-valued Datasets that are available in UCI machine learning repository are given in the bellow table.

Table e

Name of the Dataset	Number of classes	Number of normal attributes	Number of multi - valued attributes	Number of binary attributes
“Promoter”	106	2	57	-
“Hayes-Roth”	160	3	4	-
“Breast cancer”	286	2	7	3
“Monks-1”	432	2	4	2
“Monks-2”	432	2	4	2
“Monks-3”	432	2	4	2
“Balance”	625	3	4	-
“Soya large”	683	19	19	16
“Tic-tac-toe”	958	2	9	-
“Car”	1728	4	6	-
“DNA”	3190	3	60	-
“Mushroom”	8124	2	18	4
“Nursery”	12960	2	7	1

V. Similarity Measures for Multi-Valued Data Sets

In general the categorization of two types of attributes in the data bases are quantitative attributes and qualitative attributes to apply data mining tasks on these databases we are needed to find the similarity measures among these type of attributes .

VI. Qualitative Type

Tversky (1977) proposed a contrast model and ratio model that is implemented by generalizing a several set of theoretical similarity models .Tversky described the objects as sets of features as a replacement of geometric points in a geometric space. To demonstrate his models, let *a* and *b* are two objects, and *X* and *Y* indicate the sets of features associated with the objects *x* and y respectively. Tversky proposed the following family of proximity measures which is called the contrast model:

$$S(x, y) = \theta\ n\ (X \cap Y)\ -\alpha\ n(X - Y)\ -\ \beta\ n(X - Y)$$

For some θ, α, $\beta \geq 0$; *n* is usually the cardinality of the set. In the previous models, the proximity between objects was determined only by their common features, or only by their distinctive features. In the contrast model, the proximity of a pair of objects is represented as a linear combination of the measures of the common and the distinctive features. The contrast model denotes proximity between objects as a weighted difference of the measures for their common and distinctive features. The given bellow family of similarity measures denotes the ratio model:

$$S(x, y) = n\ (X \cap Y)\ /\ [n(\ X \cap Y) + \alpha\ n(X - Y) + \beta\ n(Y - X)],\ \alpha,\ \beta \geq 0$$

In the ratio model, the proximity value is normalized to a value range of *0* and *1*. In Tversky's set theoretic similarity models, a feature usually denotes a value of a binary attribute or a nominal attribute but it can be extended to interval or ordinal type. For the qualitative type of multi-valued case, Tversky's set proximity can be used since we can consider this case as an attribute for an object has group feature property (e.g., a set of feature values).

VII. Attributes That are Quantitative Type

To find the proximity within the group when the attributes are multi-valued type we use group mean for the particular attribute with respect to the object by using Euclidean distance like measures but the problem with this method is it should not consider the cardinality of the elements in a group. Another approach towards this is group average which can be used to calculate inter-group proximity. In this approach group similarity is calculated by taking the average of all the inter-object measures for those pairs of objects from which each object of a pair is in diverse groups.

For example, the average dissimilarity between group P and Q can be defined as given bellow

$$D\ (P,\ Q) = \textstyle\sum_{i=1}^{n} d(p,q)/n),$$

where *n* is the cardinality of object-pairs, *d* (*a*, *b*) is the variation metric for the i^{th} pair of objects p and q where p∊ *P* And q ∊ Q. In calculating group similarity using on group average, decision on whether we compute the average for every probable pair of similarity or the average for a subset of possible pairs of similarity is required.

VIII. Algorithms for Clustering

This paper is going to present different algorithms for clustering by considering the properties of multi-valued Data characteristics like formation, noise, dimensionality of the attributes, algorithm implementation, dimension of the data, and shape of the cluster. The overview of algorithms for clustering is given bellow.

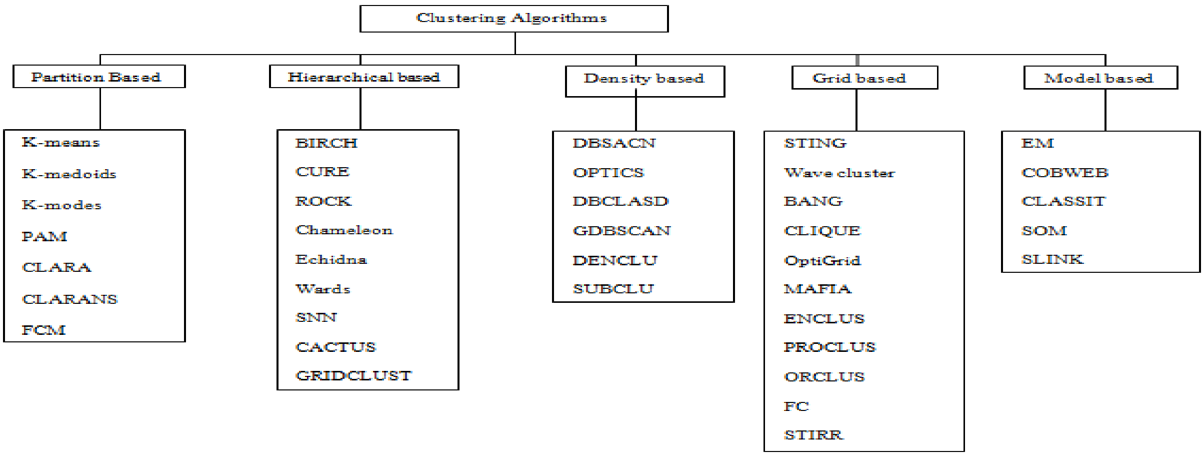


Figure 1

IX. Partitioning Algorithms for Clustering

The data objects in these types of methods are initially considered as a single cluster and this can be partitioned into different clusters depending on the number of clusters required. The objects are assigned into partitions by iteratively placing the points between the partitions. There are different re arrangement schemes that iteratively reassign the points among the pre defined number of clusters not like hierarchical methods clusters not allowed revisiting the previously formed clusters. These algorithms Progressively develop the quality of clusters.

In all of the partitioning algorithms the number of clusters is decided initially, if we want to find the number of clusters automatically we can think data comes from the mixture of various probability distributions .the major advantage of probabilistic approach is interpretability of the clusters that are formed. The summarized cluster representation can be done by allowing the measures of intra cluster similarity, an another approach to measure the quality of clustering is the definition of objective function . in general partitioning algorithms constructs the clusters in non-convex shape.

Some partitioning algorithms for clustering are K_means, K_modes, K_medoids (PAM, CLARA, CLARANS, and FCM). Partition based algorithms can found clusters of Non convex shapes. The pros and cons of partitioning clustering methods are described bellow.

Advantages:

- Comparatively scalable and simple.
- These algorithms are suitable for the datasets which forms spherical shaped clusters that are well-separated

Disadvantages:

- These algorithms causes efficiency degradation in the high dimensional spaces because almost all pairs of points are at a distance of its average the reason for this is the distance between points in high dimensional spaces is not defined properly.
- The description of cluster is less explained by these algorithms.
- User needs to specify the number of clusters in advance.
- More sensitive to initialization phase that means final results depends on the clusters that are formed at initial, also sensitive to noise and outliers.
- These are in efficient to deal with non-convex shaped clusters of unreliable size and density.

X. Hierarchical Algorithms for Clustering

Hierarchical clustering algorithms are of two types they are Agglomerative (top-bottom) and Divisive (bottom- top). In Agglomerative Hierarchical clustering initially one object is selected and the other objects are embedded consecutively to form bigger clusters the merging of objects depends on the distance like maximum, minimum and average .this procedure is repeated number of times until to form desired number of clusters. some Hierarchical algorithms for clustering are BIRCH, CURE, ROCK, Chameleon, Wards, SNN, GRIDCLUST, CACTUS in which clusters of Non convex shaped and, Arbitrary Hyper rectangular are produced.

Some times to find the proximity between sub sets it is needed to generalize the proximity between individual points. Such proximity measure is called linkage metric. This is also the case when the data base consists of multi-valued data. This type of metric impacts the Hierarchical clustering algorithms because it impacts the closeness and relation of connectivity among the attributes without considering its structure.

Mostly used inter-cluster linkage metrics are single link, average_ link, and complete link. The principal proximity of dissimilarity measure (usually distance) is computed for every pair of points with one point in the first set and another point in the second set. A specific functional operation such as minimum Average or maximum is applied to pair-wise proximity of dissimilarity measures an example for such functional operation is given bellow:

$$D\ (C_1, C_2\ )=\text{functional operation}\ \{d\ (x,y)/\ x\in C_1,\ y\in C_1\}$$

The pros and cons of partitioning clustering methods are described bellow

Advantages:

- These algorithms are flexible with respect to the level of granularity.
- similarity or distance proximities are easily managed.
- These algorithms can be applicable to any types of attributes.

Disadvantages:

- It is difficult to identify the termination point.
- Most of the hierarchical algorithms cannot be repeated if once constructed.
- Clusters are formed depending on the reason of their improvement.

XI. Density based Algorithms for Clustering

In density based methods the set of points which are in the space of Euclidean are partitioned into a set of points depending on the density among the points, connectivity between each pair of points and the boundary of the points that are grouped. In density based clusters each point in the cluster closely related to its adjacent neighbor, the clusters are formed in any direction that the density of the points in one cluster should be maximum. Because of this reason the cluster that depends on density are of arbitrary shaped as a result it provides a good security against outliers.

These algorithms classify the data objects into core_points, border_points and noise points to form dense regions. The clusters are formed by connecting the core points together. So there may be a chance of forming non spherical or arbitrary shaped clusters. Some well known density based algorithms for clustering are "DBSCAN", "OPTICS","DBCLASD", "GDBSCAN" ," DENCLUE" and "SUBCLUE".

XII. Advantages of Density based Algorithms for Clustering (DBSCAN)

1. All of these types' algorithms like DBSCAN are not needed to specify the number of clusters in the data in advance, which is opposite to k-means.
2. The major advantage of DBSCAN like algorithms are it finds arbitrary shaped clusters. It can even form clusters completely enclosed by (but not connected to) another different cluster. DBSCAN can perform well even though there is noise in the data and is robust to outliers.
3. The queries can be easily and speedily processed on the clusters that are formed by "DBSCAN".

XIII. Disadvantages of Density based Algorithms Clustering (dbscan)

1. DBSCAN is not entirely deterministic: Depending on the order of the data processed different types of clusters are formed, sometimes border points may belongs to more than one cluster so it can be a part of either of the cluster.
2. The quality of DBSCAN is influenced by the distance metric used in the function for finding the neighborhood in the Query. The distance metric that is used to find the distance is Euclidean distance for multi-valued data and high dimensional data.
3. If the data objects have high variation in its density then DBSCAN cannot cluster that data sets well. Since the combination of minPts and neighborhood distance ε cannot be chosen appropriately for all clusters.

XIV. Grid based Algorithms for Clustering

The density based methods needs some clarifications like density, boundary and connectivity for grouping the data objects. Another way of representation is by the consideration of attribute space. To minimize the search combination and space many different rectangular segments are considered where each segment is the Cartesian of individual different sub spaces.

The methods that partition the space into different cells or segments are called grid-based methods. Here each segment is related to a single value or a group of values (if the data is multi-valued data) is considered as a unit, as a result the concentration is turned on to the partition space. The partition is based on the characteristics of the grid that is formed from the data. The advantage of this type of arrangement of data in the form of grids is it does not depend on the order of the points that are arranged in the form of grids.

When the data is of numeric type then only density based and partition based methods give results efficiently but grid based techniques perform well for the attributes with different data types along with multi-valued data. For better performance the density based method "DENCLUE" initially uses grid structure. To structure real clusters Grid based algorithms uses subspace and hierarchical techniques for clustering. When compare to all Clustering algorithms Grid algorithms are very speedy and efficient processing algorithms. Arbitrary shaped clusters are formed by Adaptive grid algorithms such as MAFIA and AMR by using the grid cells.

XV. Model based Clustering Algorithms

A model is hypothesized for each of the clusters and tries to find Set of data points are related together based on different strategies like conceptual methods and statistical methods. For model based algorithms the well known approaches are one is neural network approach and another one is statistical approach. Algorithms such as "EM", "CLASSIT", "COBWEB SOM", and "SLINK" are some Model based clustering algorithms.

XVI. Research Challenges

We have previously discussed that the problem of clustering becomes very demanding,attractive and at the same time challenging, when the data is of type categorical attributes and multi-valued attributed. The number of algorithms for the discovery of groups in such data is restricted, confined and limited, compared to the research devoted on data sets with numerical data. Further, few algorithms deal with mixtures of values, i.e., attributes of numerical and categorical values. Ideally, a clustering algorithm should be needed to

- scale well, i.e., at most one scan of the data is necessary;
- handle deviations professionally;
- discover arbitrary-shaped clusters;
- give sensible execution times in the attendance of high dimensional data sets;
- present a concise model of the clusters;

XVII. Summary and Conclusion

In this paper, we analyzed the difficulty of generating single flat file arrangement to represent Data sets that have been generated from structured data-bases, and pointed out its inappropriate representation to represent related information, a fact that has been frequently overlooked by recent data mining investigation. To overcome these difficulties, we used a better representation scheme, called multi-valued *data set*, which allows attributes of an object to have a *set* of values, and studied how existing similarity measures for single-valued attributes could be applied to measure group similarity for multi-valued *data sets* in clustering.We also proposed a unified framework for proximity measures to work with multi-valued *data sets* with mixed types. Once the target database is grouped into clusters with similar properties, the discriminate query detection system, MASSON can find out useful characteristic information for a set of objects that fit in to a cluster. We claim that the planned representation scheme is suitable to cope with related information and that it is more communicative than the traditional single flat file format.

More prominently, the relationship information in a structured database is actually considered in clustering procedure.

References Références Referencias

1. Ribeiro, J.S., Kaufmann, K., and Kerschberg, L. (1995). Knowledge Discovery from Multiple Data bases, In Proc. Of the 1st Int'l Conf. On Knowledge Discovery and Data Mining, Quebec, Montreal.
2. Ryu, T.W and Eick, C.F. (1996a). Deriving Queries from Results using Genetic Programming, In Procee dings of the 2nd Int'l Conf. on Knowledge Discovery and Data Mining. Portland, Oregon.J. Pei, B. Jiang, X. Lin, and Y. Yuan, "Probabilistic Skylines on Uncertain Data",Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB), 2007.
3. Ryu, T.W and Eick, C.F. (1996b). MASSON: Discovering Commonalities in Collection of Objects using Genetic Programming, In Proceedings of the Genetic Programming 1996 Conference, Stanford University, San FranciscoR. Cheng, D.V. Kalashnikov, and S. Prabhakar, X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection", In Proc. Adv. Neural Inf. Process. Syst., vol. 186, p. 189, 2005.
4. Ryu ,T.W. and Eick ,C.F. (1998). Automated Discovery of Discriminate Rules for a Group of Objects in Databases, In Conference on Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA, June 11-13.R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", New York, NY, USA: Wiley, 2012.
5. Thompson, K., and Langley, P. (1991). "Concept formation in structured domains, In Concept Formation: Knowledge and Experience in Unsupervised Learning", Eds., Fisher, D.H; Pazzani, M.; and Langley, P., Morgan Kaufmann.
6. Tversky, A. (1977). "Features of similarity, Psychological review", 84(4): 327-352, July.
7. Everitt, B.S. (1993). "Cluster Analysis, Edward Arnold, Copublished by Halsted Press and Imprint", of John Wiley & Sons Inc., 3rd edition.
8. Gower, J.C. (1971). "A general coefficient of similarity and some of its properties", Biometrics 27, 857-872.
9. Koza, John R. (1990). "Genetic Programming: On the Programming of Computers by Means of Natural Selection", Cambridge, MA: The MIT Press.
10. S.B. Kotsiantis and P.E. Pintelas, "Recent Advances in Clustering: A Brief Survey",WSEAS Trans. Information Science and Applications, vol. 11, no. 1, pp. 73-81, 2004.
11. D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data", In Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 333-342, 2010.
12. Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2, 1-norm regularized discriminative feature selection for unsupervised learning",In Proc. Int. Joint Conf. Artif. Intell., Vol. 22, p. 1589, 2011.
13. J. Pei, B. Jiang, X. Lin, and Y. Yuan, "Probabilistic Skylines on Uncertain Data",Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB), 2007.
14. Y. Tao, R. Cheng, X. Xiao, W.K. Ngai, B. Kao, and S. Prabhakar, "Indexing Multi-Dimensional Uncertain Data with Arbitrary Probability Density Functions", Proc. Int'l Conf. Very Large Data Bases (VLDB), 2005.
15. R. Cheng, D.V. Kalashnikov, and S. Prabhakar, "Evaluating Probabilistic Queries over Imprecise Data", Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), 2003.
16. X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection", In Proc. Adv. Neural Inf. Process. Syst., vol. 186, p. 189, 2005.
17. F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint L2, 1-norms minimization", In Proc. Adv. Neural Inf. Process. Syst., vol. 23, pp. 1813-1821, 2010.
18. R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification", New York, NY, USA: Wiley, 2012.
19. X. Li and Y. Pang, "Deterministic column-based matrix decomposition", IEEE Trans. Knowl. Data Eng., vol. 22, no. 1, pp. 145-149, Jan. 2010.
20. W. Liu, D. Tao, and J. Liu, "Transductive component analysis", In Proc. 8th IEEE Int. Conf. Data Mining, pp. 433-442, 2008.

This page is intentionally left blank

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Security in Data Mining- A Comprehensive Survey

By Niranjan A, Nitish A, P Deepa Shenoy & Venugopal K R

University Visvesvaraya College of Engineering

Abstract- Data mining techniques, while allowing the individuals to extract hidden knowledge on one hand, introduce a number of privacy threats on the other hand. In this paper, we study some of these issues along with a detailed discussion on the applications of various data mining techniques for providing security. An efficient classification technique when used properly, would allow an user to differentiate between a phishing website and a normal website, to classify the users as normal users and criminals based on their activities on Social networks (Crime Profiling) and to prevent users from executing malicious codes by labelling them as malicious. The most important applications of Data mining is the detection of intrusions, where different Data mining techniques can be applied to effectively detect an intrusion and report in real time so that necessary actions are taken to thwart the attempts of the intruder.

Keywords : anamoly detection; classification; intrusion detection system; outlier detection; privacy preserving data minig.

GJCST-C Classification : H.2.8

Security in Data Mining- A Comprehensive Survey

Niranjan A [α], Nitish A [σ], P Deepa Shenoy [ρ] & Venugopal K R [ω]

Abstract- Data mining techniques, while allowing the individuals to extract hidden knowledge on one hand, introduce a number of privacy threats on the other hand. In this paper, we study some of these issues along with a detailed discussion on the applications of various data mining techniques for providing security. An efficient classification technique when used properly, would allow an user to differentiate between a phishing website and a normal website, to classify the users as normal users and criminals based on their activities on Social networks (Crime Profiling) and to prevent users from executing malicious codes by labelling them as malicious. The most important applications of Data mining is the detection of intrusions, where different Data mining techniques can be applied to effectively detect an intrusion and report in real time so that necessary actions are taken to thwart the attempts of the intruder. Privacy Preservation, Outlier Detection, Anomaly Detection and PhishingWebsite Classification are discussed in this paper.

Keywords: anamoly detection; classification; intrusion detection system; outlier detection; privacy preserving data minig.

I. Introduction

The term Security from the context of computers is the ability, a system must possess to protect data or information and its resources with respect to confidentiality, integrity and authenticity[1]. Confidentiality ensures that, a third party in no way would be able to read and understand the content while Integrity would not allow a third party to change or modify the content as a whole or even parts of it. Authenticity feature on the other hand would not allow a person to use, view or modify the content or the resource, if he is found to be unauthorised[2].

Those actions that compromise the availability, integrity or confidentiality of one or more resources of a computer could be termed as *Intrusion*. Preventing intrusions employing firewall and filtering router policies fail to stop these attacks. Inspite of all attempts to build secure systems, intrusions can still happen and hence they must be detected on their onset. An Intrusion detection system(IDS)[3] by employing data mining techniques can discover consistent patterns of features of a system that are useful can detect anomalies and known intrusions using a relevant set of classifiers. Using some of the basic data mining techniques such as Classification and Clustering, Intrusion can be detected easily. Classification techniques are helpful in analyzing and labelling the test data into known type of classes, while Clustering techniques are used to group objects into a set of clusters, such that all similar objects become the members of the same cluster and all other objects become members of other clusters[4]. Data mining, while allowing the extraction of hidden patterns or the underlying knowledge from large volumes of data, might pose security challenges[5]. Privacy Preserving Data Mining(PPDM)aims at safeguarding sensitive information from an un-solicited or unsanctioned disclosure[6]. A number of PPDM approaches have been proposed so far. Some of them are listed as shown in Fig. 1, based on their enforcing privacy principle.

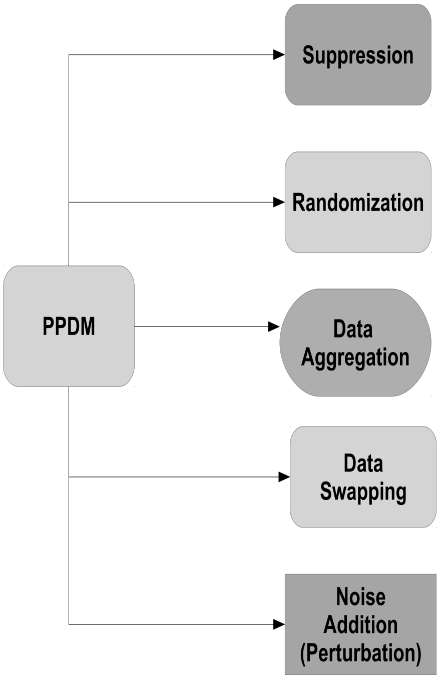


Figure 1: Privacy Preserving Data Mining Techniques

a) *Suppression*

Any private or sensitive information pertaing to an individual such as name, age, salary, address and other information is suppressed before any computation takes place. Some of the techniques employed for this suppression are Rounding(Rs/- 35462.33 may be rounded to 35,000), Generalization (Name Louis Philip

Author α σ ρ ω: Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore, India. e-mail: niranjan.a.in@ieee.org may be replaced with the initials LP and Place Hamburg may be replaced with HMG and so forth). However when data mining requires full access to sensitive values, Suppression cannot be used. An alternate way of suppression is to limit the identity linkage of a record rather than suppressing thesensitive information present within a record. This technique is referred to as *De-Identification*. *k-Anonymity* is one such de-identification technique. It ensures that protection of the data released against *Re-identification* of the persons to which the data refer[7][8]. Enforcing k-anonymity before all data are collected in one trusted place is difficult. A cryptographic solution based on Secret Sharing technique of Shamir could be used instead; this however incurs computation overhead.

b) Randomization

Assuming the presence of a central server of a company that accepts information present with many customers and performs data mining techniques for building an Aggregate Model; *Randomization* allows the customers to introduce controlled noise or randomly perturb the records and to take away true information present in it. Introduction of noise can be achieved in several ways by addition or multiplication of the values generated randomly. *Perturbation* helps Randomization technique to achieve preservation of the required privacy.

The individual records are generated by the addition of such randomly generated noise to the original data. The noise thus added to individual records cannot be recovered, resulting in the desired privacy. Randomization techniques typically involve the following steps:

1. Only after randomizing their data, the Data Providers transmit this data to the Data Receiver.
2. Data receiver computes the distribution by running a Distribution Reconstruction Algorithm.

c) Data Aggregation

Data Aggregation Techniques, in order to facilitate data analysis: combine data together from various sources. This might allow an attacker to deduce private and invidual-level data and to identify the party. When the extracted data allows the data miner to identify specific individuals, his privacy is considered to be under a serious threat. To prevent data from being identified, it may be anonymized immediately after the aggregation process. However, the Anonymized data sets can still contain enough information that could be used for the identification of individuals[9].

d) Data Swapping

Data swapping process involves swapping of values across different records for the sake of privacy-preservation. Without perturbing the lower order totals of the data, privacy of data can still be preserved allowing aggregate computations to be performed exactly as before. Since this technique does not follow randomization, it can be used in conjunction with other frameworks such as k-anonymity without violating the privacy definitions for that model.

e) Noise Addition/Perturbation

Differential privacy through the addition of controlled noise provides a mechanism that maximizes the accuracy of queries while minimizing the chances of identification of its records[10]. Some of the techniques used in this regard are:

1. Laplace Mechanism
2. Sequential Composition
3. Parallel Composition

The rest of this paper is structured as follows: Section- II covers a brief review of Classification and Detection of intrusions by employing various Data Mining Techniques, while Clustering techniques and their applications in Intrusion Detection are presented in Section-III. PPDM techniques and their necessity along with various types of PPDM are discussed in Section-IV. An overview of Intusion Detection System is discussed in Section-V. Phishing Website Classification using Data Mining Techniques are presented in Section-VI. Artificial Neural Networks(ANN) are presented in Section-VII. Section- VIII presents Anomaly Detection/Outlier Detection. Section- IX describes the various ways of Mitigating Code Injection Attacks.

II. Classification and Detection Using Data mining Techniques

Malware computer programs that replicate themselves in order to spread from one computer to another computer are called as worms. Malware includes worms, computer viruses, Trojan Horse, key loggers, adware, spyware Port scan worm, UDP worm, http worm, User to Root Worm and Remote to Local Worm and other malicious code[11]. Attackers write these programs for various reasons varying from interruption of a computer process, gathering sensitive information, or gaining entry to private systems. Detecting a worm on the internet is very important, because it creates vulnerable points and reduces the performance of the system. Hence it is essential to detect the worm on the onset and classify it using data mining classification algorithms much before it causes any damage.

Some of the classification algorithms that can be used are Random Forest, Decision Tree, Bayesian and others[12]. A majority of worm detection techniques use Intrusion Detection System(IDS) as the underlying principle. Automatic detection is challenging because it is tough to predict what form the next worm will take. IDS can be classified into two types namely Network based IDS and Host based IDS. The Network based Intusion Detection System reflects network packets before they spread to an end-host, while the Host based Intusion Detection System reflects network packets that are already spread to the end-host. Moreover, the Host based detection studies encode network packets so that the stroke of the internet worm may be struck. When we focus on the network packet without encoding, we must study the performances of traffic in the network. Several machine learning techniques have been used in the field of intrusion and worm detection systems. Thus, Data Mining and in particular Machine Learning Technique has an important role and is essential in worm detection systems. Using various Data Mining schemes several new techniques to build several Intrusion Detection models have been proposed. Decision Trees and Genetic Algorithms of Machine Learning can be emoloyed to learn anomalous and normal patterns from the training set and classifiers are then generated based on the test data to label them as Normal orAbnormal classes. The data that is labelled as Abnormal could be a pointer to the presence of an intrusion.

a) Decision Trees

Quinlan's decision tree technique, is one of most popular machine learning techniques. The tree is constructed using a number of decision and leaf nodes following divide-andconquer technique[12]. Each decision node tests a condition on one of the attributes of the input data and can essentially have a number of branches, to handle a separate outcome of the test. The result of decision may be represented as a leaf node. A training data set *T* is a set of n-classes {*C1, C2 ,..., Cn*}. *T* is treated as a leaf when it comprises of cases belonging to a single class. If *T* is empty with no cases, it is still treated a leaf and the major class of the parent node is given the related class. A test based on an attribute ai of the training data is performed when *T* consists of multiple classes, *T* is split into *k* subsets {T1, T2, ..., Tk}, where *k* gives the number of test outcomes. The process is recursed over each Tj, where 1 <= j<= n, until every subset belongs to a single class. Choosing the best attribute for each decision node while constructing the decision tree is very crucial. The C4.5-DT adopts Gain Ratio Criterion for the same. According to this criterion, an attribute that provides maximum information gain and that reduces the bias in favor of tests is chosen. The tree thus built can then be used to classify the test data, whose features are same as that of the training data. The test is carried out starting from the root node. Based on the outcome, one of the branches leading to a child is followed. As long as the child is not a leaf, the process is repeated recursively. The class and its corresponding leaf node is given to the test case being examined.

b) Genetic Algorithms(GA)

A machine learning approach of solving problems by employing biological evolution techniques are called *Genetic Algorithms(GA)*. They can be effectively used to optimize a population of candidate solutions. GA makes use of data structures that are modelled on chromosomes and they are subjected to evolution using genetic operators namely: selection, crossover and mutation[13]. Random generation of a population of chromosomes is performed in the beginning. The population thus formed comprises of all possible solutions of a problem and are considered the candidate solutions. Different positions of a chromosome called '*genes*' are encoded as bits, characters or numbers. A function called *Fitness Function* evaluates the goodness of each chromosome based on the desired solution. *Crossover* operator simulates natural reproduction while *Mutation* operator simulates mutation of the species. The *Selection* operator chooses the fittest chromosomes[14]. Fig 2. depicts the operations of Genetic Algorithms. Before using GA for solving various problems, following three factors have to be considered

1. Fitness function
2. Individuals representation and
3. Parameters of GA

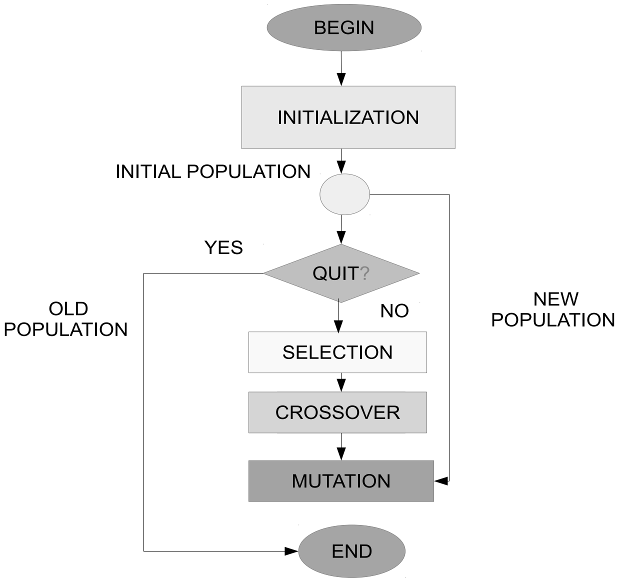


Figure 2: Flowchart for a GA

GA based approach can be incorporated for designing Artificial Immune Systems. Using this approach, Bin et al.,[15] have proposed a method for smartphone malware detection where static and dynamic signatures of malwares are extracted and malicious scores of tested samples are obtained.

c) Random Forest

A classification algorithm that is made up of a collection of tree structured classifiers, and that chooses the winner class based on the votes casted by the individual trees present in the forest is called the *Random Forest Algorithm*. Each tree is constructed by picking up random data from a training dataset. The chosen dataset may be split up into training and testsets. The major chunk of the dataset goes into the training set while the minor chunk forms the test set. The tree construction involves the following steps:

1. If the training set has *N* cases, a sample of *N* cases is randomly selected from the original dataset. This sample corresponds to training set that is used for growing the tree.
2. *m* variables out of the *M* input variables are chosen randomly, the node is split based on the best split on this *m* value. *m* is held constant while growing the forest.
3. Each tree in the forest is grown to the largest possible extent. No Trimming or Pruning is performed on the individual trees.
4. All classification trees thus formed are combined to form the random forest. Since it can fix problem of overfitting on large dataset and can train/test rapidly on complex data set, it is sometimes referred to as *Operational Data mining* technique.

Each classification tree is exclusive and is voted for a class. Finally, a solution class is constructed based on the maximum votes assigned.

d) Association Rule Mining (ARM)

Association-rule mining discovers interesting relations between a set of attributes in datasets[16]. The datasets and their inter-relationship can be represented as association rules. This information can be used for making strategic decisions about different activities such as, promotional pricing, shelf management and so on[17]. Traditional Association rule mining involves a data analyst being given datasets of different companies for the purpose of discovering patterns or association rules that exist between the datsets[18]. Although, we can achieve sophisticated analysis on these extremely large datasets in a cost-effective manner[19], it poses security risk[20] for the data owner whose sensitive information can be deduced by the dataminer[21]. Even today, association rule mining is one of the widely used pattern discovery methods in KDD.

Solving an ARM problem basically involves traversing the items in a database, which can be done using various algorithms based on the requirement[22]. ARM algorithms are primarily categorised into BFS (Breadth First Search) and DFS (Depth First Search) methods based on the strategy used to traverse the search space[23]. The BFS and DFS methods are further classified into Counting and Intersecting, based on how the support values for the itemsets are determined. The algorithms Apriori, Apriori-TID and Apriori-DIC are based on BFS with Counting strategies, while the Partition algorithm is based on BFS with Intersecting strategies. The FP-Growth algorithm on the otherhand, is based on DFS with Counting strategies while ECLAT is based on DFS with Intersecting[24][25]. These algorithms can be optimized specifically for improving the speedup [26][27].

BFS with Counting Occurences: The common algorithm in this category is the Apriori algorithm. It utilizes the downward closure property of an itemset, by pruning the candidates with infrequent subsets before counting their supports.The two metrics to be considered while evaluating the association rules are: *support* and *confidence*. BFS offers the desired optimization by knowing the support values of all subsets of the candidates in advance. The limitation of this approach is increased computational complexity in rule extraction from a large database. Fast Distributed Mining(FDM) algorithm is a modified, distributed and unsecured version of the Apriori algorithm[28]. The advancements in data mining techniques, have enabled organizations in using data more efficiently.

In Apriori, the candidates of a cardinality *k* are counted by a single scan of the entire database. Looking up for the candidates in each transaction forms the most crucial part of the Apriori Algorithm. For this purpose, a hashtree structure is used[29]. Apriori-TID an extension of Apriori, represents each transaction based on the current candidates it contains, unlike normal Apriori that relies on raw database. Apriori-Hybrid combines the benefits of both Apriori and Apriori-TID. Apriori-DIC another variation of Apriori, tries to soften the separation that exists between the processes, counting and candidate generation. This is done by using a prefix-tree.

BFS with Intersections: A Partition Algorithm is similar to the Apriori algorithm that uses intersections rather than counting occurences for the determination of support values. The partitioning of itemsets could result in the exponential growth of intermediate results beyond the physical memory limitations. This problem can be overcome, by splitting the database up into a number of chunks that are smaller in size and each chunk is treated independently. The size of a chunk is determined such that all intermediate lists can fit into memory. An additional scan can optionally be performed to ensure that the itemsets are not only locally frequent but also are globally frequent.

DFS with Counting Occurences: In Counting, a database scan for each reasonable sized candidate set is performed. Because of the involvement of computational overhead in database scanning, the simple combination of DFS and Counting Occurences is practically irrelevant. FP-Growth on the otherhand uses a highly compressed representation of transaction data called *FP-Tree*. An FP-Tree is generated by counting occurences and performing DFS.

DFS with Intersections: The algorithm *ECLAT* combines DFS with the list intersections to select agreeable values. It makes use of an optimization technique called *Fast Intersections*. It does not involve the process of splitting up of the database since complete path of classes beginning from the root would be maintained in the memory. As this method eliminates most of the computational overhead the process of mining association rules becomes faster.

III. Clustering

Clustering is one of the widely used discovery methods in data mining. It allows to group a set of data in such a way that, Intra-Cluster similarity are maximized while minimizing the Inter-Cluster similarity are minimized. Clustering involves unsupervised learning of a number of classes that are not known in advance. The clustering algorithms can be broadly clasified into the following types and are listed in Fig.3

1. Connection Based or Hierarchical Clustering
2. Centroid Based
3. Distribution Based
4. Density Based
5. Recent Clustering Techniques and
6. Other Clustering Techniques

a) *Connection Based Clustering*

Connection Based (Hierarchical) clustering, is based on the idea of objects being more related to closer objects than to the distant objects. The Connection Based Clustering algorithms consider the distance between the objects to connect them to form *clusters*. These algorithms provide an extensive hierarchy of merging clusters at particular distances, instead of single partitioning of dataset. A *Dendrogram* is used to represent clusters. Its *y-axis* shows the merging distance of the clusters and the *x-axis*, for placing the objects, ensuring that the clusters do not mix. There are various types of Connection based clustering based on the way the distances are computed such as: Single-Linkage Clustering that involves determining of the minimum of object distances, Complete-Linkage Clustering where the maximum of object distances is computed and Unweighted Pair Group Method with Arithmetic Mean (UPGMA), or Average Linkage Clustering. Selecting appropriate clusters from the available hierarchy of clusters, could be achieved either using Agglomerative or Divisive Clustering.In Agglomerative Clustering, we begin with single objects and conglomerate them into clusters while in Divisive clustering, we start with the complete data set and isolate it into segments.

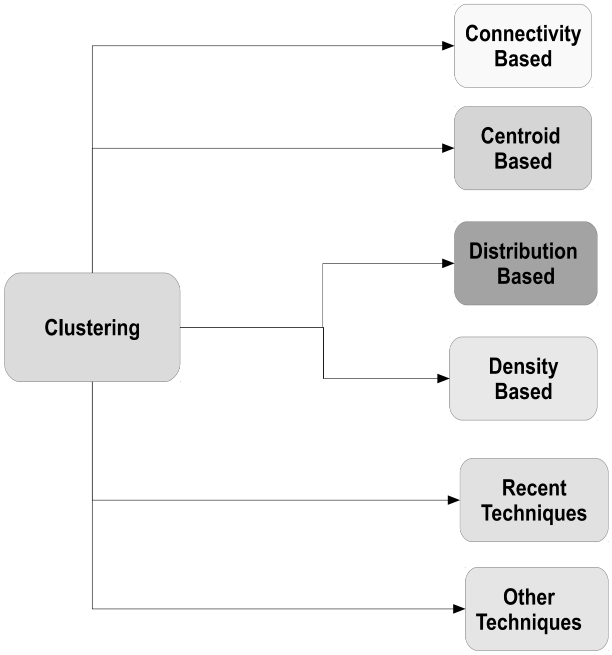


Figure 3: Types of Clustering

b) *Centroid Based Clustering*

Centroid-based clustering may have clusters that are represented by a vector, which necessarily is not a member of the data set or may have clusters strictly restricted to the members of the dataset. In *k-means Clustering* algorithm, the number of clusters is limited to size *k*, it is required to determine *k* cluster centers and assigning objects to their nearest centers.

The algorithm is run multiple times with different *k* random initializations to choose the best of multiple runs[30]. In *kmedoid clustering*, the clusters are strictly restricted to the members of the dataset while in *k-medians clustering*, only the medians are chosen to form a cluster. The main disadvantage of these techniques is that the number of clusters *k* is selected beforehand. Furthermore, they result in incorrectly cut borders in between the clusters.

c) *Distribution Based Clustering*

Distribution-based clustering technique forms clusters by choosing objects that belong more likely to the same distribution. One of the most commonly preferred distribution techniques is the Gaussian Distribution. It suffers from the overfitting problem where a model cannot fit into set of training data.

d) *Density Based Clustering*

In this type of clustering, an area that is having higher density than the rest of the data set is considered as a cluster. Objects in the sparse areas are considered to be noise and border points. There are three commonly used Density-based Clustering techniques namely: DBSCAN, OPTICS and Mean- Shift. DBSCAN is based on connecting points that satisfy a density criterion within certain distance thresholds. The cluster thus formed may consist of all density-connected objects and objects that are within these objects range free to have an arbitrary shape.

e) *Recent Clustering Techniques*

All the standard clustering techniques fail for highdimensional data and so some of the new techniques are being explored. These techniques fall into two categories namely: Subspace Clustering and Correlation Clustering. In Subspace Clustering, the clustering model specifies a small list of attributes that should be considered for the formation of a cluster while in Correlaton Clustering,the model along with this list of attributes it also provides the correlation between the chosen attributes.

f) *Other Techniques*

One of the most basic clustering techniques is the BSAS(Basic Sequential Algorithmic Scheme). Given the distance *d(p, C)* between a vector point *p* and a cluster *C*, the maximum number of clusters allowed *q* and threshold of dissimilarity 0, the BSAS constructs the clusters even when the number of clusters to be formed is not known in advance.

Every newly presented vector is either assigned to an already existing cluster or a new cluster is created, depending on the distance to the already present clusters.

g) *Clustering applications in IDS*

Clustering technique may be effectively used in the process of Intrusion Detection. The setup is depicted in Fig. 4. Alerts generated by multiple IDSs belonging to both Network and Host types are logged into a centralized database. The alert messages arriving from diffrent IDSs will be in different formats. Before passing them into the server, a preprocessing step is needed to bring them all into some uniform format [31].

Best effort values are chosen for the missing attributes during the preprocessing stage. The timestamp information may have to be converted into seconds for the sake of comparison. Different IDSs may use different conventions for naming a single event and hence it is required to standardize the messages. Each alert may be added with an unique ID to keep track of the alerts. After preprocessing and normalizing alerts they are passed to the first phase to perform filtering and labeling functions. To minimise the number of Alerts, it is a good idea to employ Alert Fusion during which alerts with same attributes that differ by a small amount of time are fused together. Alert Fusion makes the generalization process fast. Generalization involves the addition of hierarchical background knowledge into each attribute. On every iteration of this process, the selected attribute is generalized to the next higher level of hierarchy and those alerts which have become similar by now are grouped together.

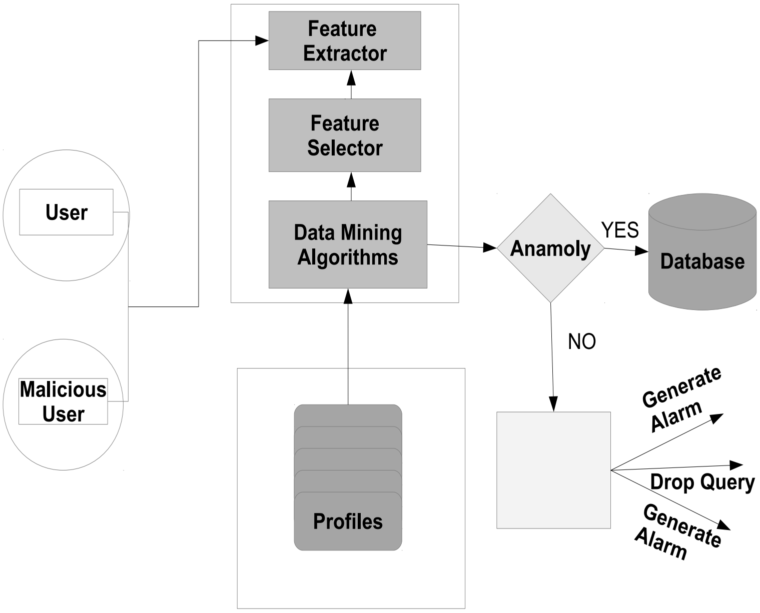


Figure 4: Use of Clustering in IDS

IV. Privacy Preserving Data Mining (PPDM)

Privacy Preserving Data Mining techniques aim at the extraction of relevant knowledge from large volumes of data while protecting any sensitive information present in it. It ensures the protection of sensitive data to conserve privacy and still allowing us to perform all data mining operations efficiently. The two types of privacy concerned data mining techniques are:

1. Data privacy
2. Information privacy

Data privacy focuses on the modification of the database for the protection of sensitive data of the individuals while Information privacy focuses on the modification for the protection of sensitive knowledge that can be deduced from the database.

Alternatively we can say that Data privacy is corcerned about providing privacy to the input while Information privacy on the otherhand is about providing privacy to the output. Preserving personal information from revealation is the main focus of a PPDM algorithm[32]. The PPDM algorithms rely on analysing the mining algorithms for any side effects that are acquired during Data privacy. The objective of Privacy Preserving Data Mining is building algorithms that transform the original data in some mannner, so that both the private data and knowledge are not revealed even after a successful mining process. Only when some relevant adequate benefit is found resulting from the access, the privacy laws would allow the access.

Multiple parties may sometimes wish to share private data resulting after a successful aggregation[33] without disclosing any sensitive information from their end[34]. Consider for example, different Book stores with respective sales data that is in a way considered to be highly sensitive, may wish to exchange partial information among themselves to arrive at the aggregate trends without disclosing their individual store trends. This requires the use of secure protocols for sharing the information across multiple parties. Privacy in such cases should be achieved with high levels of accuracy[35].

The data mining technology by principle is neutral in terms of privacy[36]. The motive for which a data mining algorithm is used could either be good or malicious[37]. Data mining has expanded the investigation possibilities[38] to enable researchers to exploit immense datasets on one hand[39], while the malicious use of these techniques on the other hand has introduced threats of serious nature against protection of privacy[40].

Table 1: Research Progress in PPDM

Authors	Algorithm	Performance	Future enhancement
Boutet et al.(2015)[45]	kNN	Better than Randomization scheme	Can consider all attacking models
Tianqing et al.(2015)[46]	Correlated Differential Privacy (CDP)	Enhances the utility while answering a large group of queries on correlated datasets	Can be experimented with Complex Applications
Bharath et al.(2015)[47]	PP k-NN classifier	Irrespective of the values of k, it is observed that SRkNNo is around 33% faster than SRkNN. E.g., when k=10, the computation costs of SRkNNo and SRkNN are 84.47 and 127.72 minutes, respectively (boosting the online running time of Stage 1 by 33.86%)	Parallelization is not used
Nethravathi et al.(2015)[48]	PPDM	Reduced misplacement clustering error and removal of data that is sensitive and correlated	Works only for numerical data
Mohammed et al.(2014)[49]	Differential Privacy	More secured under the Semi-Honest model	Overcoming Privacy Attack
Vaidya et al.(2014)[50]	Distributed RDT	Lower Computation and Communication cost	Limited information that is still revealed must be checked
Lee(2014)[51]	Perturbation methods	Capable of performing RFM Analysis	Partial disclosure is still possible

Discovering the base of privacy preserving data mining algorithms and connected privacy techniues is the need of the hour[41]. We are required to answer few questions in this regard such as

1. Evaluation of these algorithms with respect to one another
2. Should privacy preserving techniques be applied to each of the data mining algorithms? Or for all applications?
3. Expanding the places of usage of these techniques.
4. Investigating their use in the fields of Defense and Intelligence, Inspection and Geo-Spatial applications.
5. The techniques of combining confidentiality, privacy and trust with high opinion to data mining.

To answer these questions, research progresses in both data mining and privacy are required. Proper planning towards developing flexible systems is essential[42]. Few applications may demand *pure data mining* techniques while few others may demand *privacy-preserving data mining*[43]. Hence we require flexible techniques in data mining that can cater to the the changing needs[44]. The research progress made so far in the area of PPDM is listed in Table 1.

Distributed Privacy Preserving Data Mining(DPPDM):

The tremendous growth of internet in the recent times is creating new opportunities for distributed data mining[52], in which, mining operations performed jointly using their private inputs[53]. Often occurence of mining operations between untrusted parties or competitors, result in privacy leakage[54]. Thus, Distributed Privacy Preserving Data Mining(DPPDM)[10][55] algorithms require a high level of collaboration between parties to deduce the results or to share mining results that are not sensitive. This could sometimes result in the disclosure of sensitive information.

Distributed data mining are classified as Horizontally Partitioned Data and Vertically Partitioned Data. In a Horizontally partitioned data framework, each site maintains complete information on an unique set of entities, and the integrated dataset consists of the union of all of these datasets. Vertically Partitioned Data framework on the otherhand involves each site, maintaining different types of information and each dataset and has only limited information about same set of entities.

Privacy feature can limit the information leakage caused by the distributed computation techniques[56].

Each non-trusting party can compute its own functions for unique set of inputs, revealing only the defined outputs of the functions. Apart from hiding sensitive information, the privacy service also controls the information and its uses by involving various number of negotiations and tradeoffs between hiding and sharing.

All efficient PPDM algorithms are based on the assumption that it is acceptable to release the intermediate results obtained during the data mining operations. Encryption techniques solve the data privacy problem and their use would make it easy to perform data mining tasks among mutual untrustworthy parties, or between competitors. Due to its privacy concern, Distributed Data Mining Algorithms employ encryption techniques. Encryption is used in both approaches(horizontally and vertically partitioned data) of Distributed Data mining without much stress on the effiency of encryption technique used.

If the data are stored on different machines and partitioning is done row-wise, it is called horizontal partitioning and if the data are stored and partitioned column wise then it is called vertical partitioning. An overview of the same is depicted in Fig.5.

The objective of data mining techniques is to generate high level rules or summaries and generalize across populations, rather than revealing information about individuals but they work by evaluating individual data that is subject to privacy concerns. Since much of this information held by various organizations has already been collected, providing privacy is a big chalenge. To prevent any correlation of this information, control and individual safeguards must be separated to be able to provide acceptable privacy. Unfortunately, this separation makes it difficult to use the information for the identification of criminal activities and other purposes that would benefit the society. Proposals to share information across agencies to combat terrorism and other criminal activities, would also remove the safeguards imposed by separation.

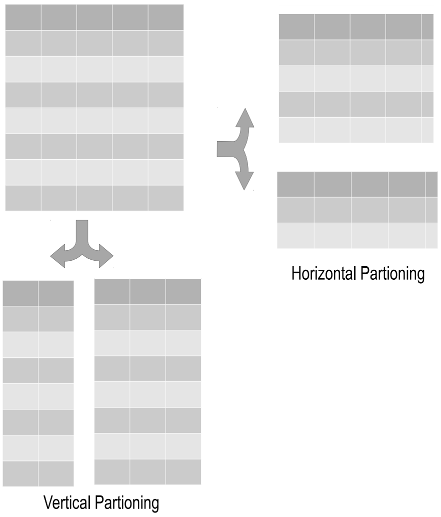


Figure 5: Horizontal and Vertical Partioning Techniques

Many of the complex socio-technical systems suffer from an inadequate risk model that focuses on the use of Fair Information Practice Principles(FIPPs). Anonymization suffers from the risk of failure, since the circumstances surrounding its selection are ignored. A Hybrid approach that combines privacy risk model with an integrated anonymization framework involving anonymization as the primary privacy risk control measure can be considered instead[57].

Public-Key Program Obfuscation: The process of making a program uncomprehensible without altering its functionality is called Program Obfuscation. A program that is obfuscated should be a *virtual black box* meaning, if it is possible for one to compute something from it, it should also be possible to compute the same even from the input-output behavior of the program. Single-Database Private Information Retrieval can be considered a type of public-key program obfuscation. Given a program *p* from a class of programs *C*, and a security parameter *s*, a public-key program obfuscation function compiles *p* into *(P, Dec)*, where *P* on any input computes an encryption of what *p* would compute on the same input and the decryption algorithm *Dec* decrypts the output of *P*. That is, for any input *i*, $Dec(P(i)) = p(i)$, but for given code *P* it is impossible to distinguish which *p* from the class *C* was used to produce *P*. The program encoding length $|P|$ must depend only on $|p|$ and *s*, and the output length of $|P(i)|$ must polynomially depend only on $|p(i)|$ and *k*.

Secure Multi-party Computation: Distributed computing involves a number of distinct,and connected computing devices that wish to carry out a combined computation of some function. For example, servers holding a distributed database system, may wish to update their database. The objective of secure multiparty computation is to allow parties to carry out distributed computing tasks in a secure way[33]. It typically involves the parties carrying out a computation based on their private inputs and neither of them willing to disclose its own input to other parties. The problem is conducting such a computation by preserving the privacy of their inputs. This problem is called the Secure Multi-party Computation problem (SMC)[34]. Consider the problem of two-parties who wish to securely compute the median. The two parties have with them two separate input sets *X* and *Y*. The parties are required to jointly compute the median of the union of their sets *X U Y*, without revealing anything about each other's set. Association Rules can be computed in an environment where different information holders have different types of information about a common set of entities.

V. Intrusion Detection System(IDS)

Intrusion detection systems aim at the detection of an intrusion on its onset[58]. A high level of human expertise and significant amount of time are required for the development of a comprehensive IDS[59]. However, IDSs that are based on the Data Mining techniques require less expertize and yet they perform better. An Intrusion Detection System detects network attacks against services that are vulnerable [60], attacks that are data driven on applications, privilege escalation[61], logins that are un-authorized and access to files that are sensitive in nature[62]. The data mining process also efficiently detects malware from the code[63], which can be used as a tool for cyber security[64][65]. An overview of an Intrusion Detection System is presented in Fig 6.

An IDS is basically composed of several components such as, sensors, a console monitor and a central engine[66]. Sensors generate security events while all events and alerts are monitored and controlled by the Console Monitor and the Central Engine records events in a database and generate alerts based on a set of rules[67]. An Intrusion detection system[68] can be classified depending on the location and the type of Sensors and based on the technique used by the Central engine for the generation of alerts. A majority of IDS implementations, involve all of the three components integrated into a single device.

Current virus scanner methodology makes use of two parts namely a Detector based on signatures and a Classifier based on the heuristic rules for the detection of new viruses. The signature-based detection algorithms rely on signatures that are unique strings of known malicious executables for the generation of detection models.The disadvantages of this approach are: it is more time-consuming and fails in detecting new malicious executables. Heuristic classifiers on the other hand are generated by a set of virus experts for the detection of new malicious executables.

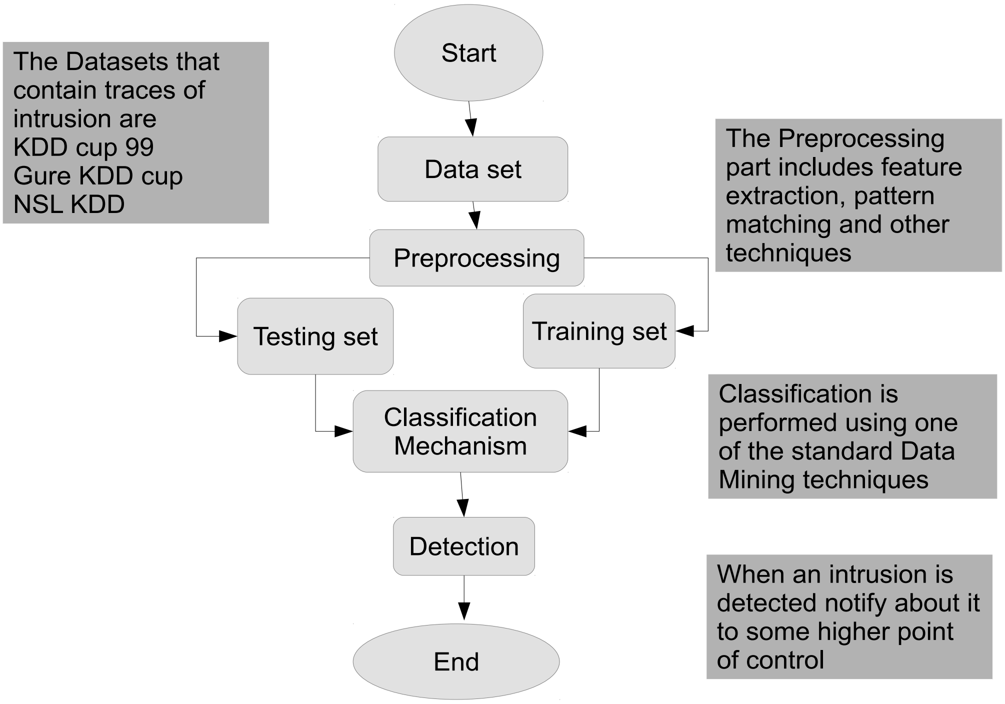

The Datasets that contain traces of intrusion are
KDD cup 99
Gure KDD cup
NSL KDD
Start
Data set
Preprocessing
Testing set
Training set
Classification Mechanism
Detection
End
The Preprocessing part includes feature extraction, pattern matching and other techniques
Classification is performed using one of the standard Data Mining techniques
When an intrusion is detected notify about it to some higher point of control

Figure 6: An overview of an Intrusion Detection System

a) Types of IDS

An intrusion could be detected either on a individual system or on a network and accordingly we have three types of IDS namely: Network Based, Host Based and Hybrid IDS.

i. *Network Based IDS*

Because of their increasingly vital roles in modern societies, computer networks have been targeted by enemies and criminals. For the protection of our systems, it is very essential to find the best possible solutions. Intrusion prevention techniques such as, authentication technqiues involving passwords or biometrics[69], programming errors avoidance, and protection of information using encryption techniques have been widely used as a first line of defense. Intrusion prevention techniques as the sole defense mechanism are not sufficient enough to combat attacks. Hence, it can therefore be used only as a second line of defense for the protection of computer systems[70].

An Intrusion Detection system must protect resources such as accounts of users[71], their file systems and the system kernels of a target system and must be able enough to characterize the legitimate or normal behavior of these resources by involving techniques that compare the ongoing system activities with already established models and to identify those activities that are intrusive[72][73]. Network packets are the data source for Network-Based Intrusion Detection Systems. The *NIDS* makes use of a network adapter to listen to and analyse network traffic as the packets travel across the network. A Network based IDS generates alerts upon detecting an intrusion from outside the perimeter of its enterprise[74]. The network based IDSs are categorically placed at strategic points on LAN to observe both inbound and outbound packet[75]. Network based IDSs are placed next to the firewalls to alert about the inbound packets that may bypass the firewall[76]. Few Network-Based IDSs take custom signatures from the user security policy as input, permitting limited detection of security policy violations[77]. When packets that contain intrusion originated from authorized users, the IDS may not be able to detect[78][79].

Advantages

Some of the advantages of a Network Based IDS are as follows:

1. For enhanced security against attacks, they can be made invisible.
2. Are capable of monitoring larger networks.
3. They can function without interfering with the normal operation of a network[80].
4. It is easy to fit in an IDS into an existing network.

Disadvantages

The disadvantages are as follows:

1. Not capable enough to analyze encrypted information coming from virtual private networks.
2. Their success most of the times depend on the capabilities of the intermediate switches present in the network.
3. When the attackers fragment their packets and release them,the IDS could become unstable and crash.

ii. *Host Based IDS*

In a Host-based IDS, the monitoring sensors are placed on network resources nodes so as to monitor logs that are generated by the Host Operating System or application programs.

These Audit logs contain records of events or activities that are occuring at individual Network resources[81]. Since a Host- Based IDS is capable of detecting attacks that cannot be seen by a Network based IDS, an attacker can misuse one of trusted insiders[82]. A Host based system utilizes Signature Rule Base that is derived from security policy that is specific to a site. A Host Based IDS can overcome all the problems associated with a Network based IDS as it can alarm the security personnel with the location details of intrusion, he can take immediate action to thwart the intrusion. A Host based IDS can also monitor any unsuccessful attempts of an attacker. It can also maintain seperate records of user login and user logoff actions for the generation of audit records.

Advantages

Some of the advantages of a Host Based IDS are as follows:

1. Can detect attacks that are not detected by a Network Based IDS.
2. Operates on Operating System audit log trails, for the detection of attacks involving software integrity breaches.

Disadvantages

The disadvantages are:

1. Certain types of DoS(Denial of Service)attacks can disable them[83].
2. Not suited for detecting attacks that target the network.
3. Difficult to configure and manage every individual system.

iii. *Hybrid IDS*

Since Network and Host-based IDSs have strengths and benefits that are unique over one another, it is a good idea to combine both of these strategies into the next generation IDSs[84]. Such a combination is often referred to as a Hybrid IDS. Addition of these two components would greatly enhance resistance to few more attacks.

a. *DM techniques for IDS*

Some of the techniques and applications of data mining required for IDS include the following

1. Pattern Matching
2. Classification and
3. Feature Selection

Pattern Matching: Pattern Matching is a process of finding a particular sequence of a part of data (substring or a binary pattern), in the whole data or a packet to get a desired information[87]. Though it is fairly rigid, it is indeed simple to use. A Network Based IDS succeeds in detecting an intrusion only when the packet in question is associated with a particular service or, destined to or from a particular port. That is, only few fields of the packet such as Service, Source/Destination port address and few others have to be examined thereby reducing the amount of inspection to be done on each packet.

However, it makes it difficult for systems to deal with Trojans and their associated traffic that can be moved at will. The pattern matching can be classified into two categories based on the frequency of occurrence namely:

a) Frequent Pattern Matching and
b) Outlier Pattern Matching

a) *Frequent Pattern Matching*

These are the type of patterns which occur frequently in an audit data, i.e., the frequency of occurrence of these patterns is more compared to other patterns in the same data[82].

Determining frequent patterns in a big data helps in analyzing and forecasting of a particular characteristic of the data. For example, by analyzing the sales information of an organization, frequent pattern matching might help to predict the possible sales outcome for the future. It also helps in decision making.

The frequent pattern mining in ADAM project data is done by mining the repository for attack-free (train) data which is compared with the patterns of normal profile (train) data. A classifier is used to reduce the false positives.

b) *Outlier Pattern Matching*

Patterns that are unusual and are different from the remaining patterns and that are not noise are referred to as Outlier Patterns. Preprocessing phase eliminates noise as it is not a part of the actual data while outliers on the other hand cannot be eliminated. Outliers exhibit deviating characteristics as compared to the majority of other instances. Outliers patterns are not usual and they occur less frequently and for this reason will have minimal support in the data. These patterns can quite often point out some sort of discrepancy in data such as transactions that are fraudulent, intrusion, abnormal behavior, economy recession etc.,. The outlier pattern mining algorithms can be of two types, one that looks for patterns only at fixed time intervals, and the other that calculates monitors patterns at all times. Outlier pappers make use of special data structures such as Suffix Tree and other String Matching Algorithms.

Classification: Classification makes use of training examples for learning a model and to classify samples of data into known classes[88]. A wide range of classification techniques ranging from Neural Networks, Decision Trees, Bayesian classifier[89], Bayesian Belief Networks and others are used in applications that involve Data Mining techniques. Classification typically involves steps that are outlined below:

Table 2: Research Progress in IDS

Authors	Algorithm	Performance	Future enhancement
M Vittapu et al.(2015)[85]	SVM Classification	TPR of 96% and FPR of 5%	Can be experimented with other techniques
Mitchell et al.(2015)[61]	Behavior Rule Analysis	Better performance	Can be tested with other techniques
Jabez J et al.(2014)[98]	Hyperboli Hopfiel Neural Network(HHNN)	Detection rate of about 90%	Can be improved
S Abadeh et al.(2014)[151]	Genetic Fuzzy System	Best tradeoff in terms of the mean F-measure,the average accuracy and the false alarm rate	A Multi-objective Evolutionary Algorithm for maximizing performance metrics may be considered
Soni et al.(2014)[86]	Feature Selection	Better classification	Can consider NSL-KDD

1. Creation of a training dataset
2. Identification of classes and attributes
3. Identification of attributes that are useful for classification
4. Relevance analysis
5. Learning the Model using training examples
6. Training the set
7. Using the model for the classification of unknown data samples.

Bayesian Classifiers: The Naive Bayesian approach assumes, the attributes to be independent in condition. Although it works under this assumption, the Naive Bayesian classifiers yield results that are satisfactory because they focus on identifying the classes for the instances instead of their probabilities. Spam Mail classification and Text classification applications extensively use Naive Bayesian classifiers for they are less error prone. However, their disadvantage is that they require probabilities in advance. The probability information that is required by them is extremely huge which consist number of classes, their attributes and the maximum cardinality of attributes. The space and computational complexity of these classifiers increase exponentially.

Support Vector Machine(SVM): Support Vector Machine is one of the learning methods extensively used for the Classification and Regression analysis of Linear and Non-linear data[90]. It maps input feature vectors into a higher dimensional space using non-linear mapping techniques. In SVM, the classifier is created by the linear separation of hyperpalnes and linear separation is achieved using a function called kernel.The Kernel transforms a linear problem by mapping it into feature spaces.

Some of the commonly used kernel functions are Radial basis, sigmoid neural nets and polynomials. Users specify one of these functions while training the classifier and it selects support vectors along the surface of this function. The SVM implementation tries to achieve maximum separation between the classes[91]. Intrusion detection system involves two phases namely training and testing. SVMs are capable of learning a larger set of patterns and can provide better classification, because the categorizing complexity is independent of the feature space dimensionality[92]. SVMs can update the training patterns dynamically with the availability of new pattern during classification. For the efficient classification it is required to reduce the dimensionality of the dataset. To do this we have *Feature Selection*.

iii. *Feature Selection(FS)*

The process of reducing the dataset dimensionality by selecting a subset of the features from the given set of features is called Feature Selection[93]. FS involves discarding of redundant and irrelevant features. FS is considered to be an efficient machine learning technique that helps in building classification systems which are efficient. With the reduction in subset dimensionality, the time complexity is reduced with improved accuracy, of a classifier. Information Gain is a proposition of feature selection that can be used to compute entropy cost of each attribute. An entropy cost can be called as a rank. Rank of each feature represents its importance or association with an solution class that is used to recognize the data. So a feature with comparatively higher rank will be one of the most important features for classification. The three standard approaches that are commonly followed for feature selection are embedded technique, filter technique, and wrapper technique.

FS runs as a part of data mining algorithms, in Embbedded technique. Feature selection is independent of the classifier used in case of Filter method, while in Wrapper method features are chosen specifically to the intended classifier. Filter method uses an arbitrary statistical way for the selection of features whereas wrapper method uses a learning algorithm to find the best subset of features. Wrapper approach is more expensive and requires more computational time than the filter approach but gives more accurate results compared to filter technique.

VI. Phishing Websites Classification

In the art of emulating a website of a trusted and creditable firm with the intention of grabbing users' private information (ussername, password) is called *phishing*. Fake websites areususlly created by dishonest people to masquerade honest websites. Users unknowingly lose money due to phishing activities of attackers. Online trading therefore demands protection from these attacks and is considered a critical step. The prediction and classification accuracy of a website depends on the goodness of the extracted features. Most of the internetusers feel safe against phishing attacks by utilizing antiphishing tool, and hence the anti-phishing tools are required to be accurate in predicting phishing[94]. Phishing websites give us a set of clues within its content parts and through security indicators of the browsers[95]. A variety of solutions have been proposed to tackle the problem of phishing. Data mining techniques involving Rule based classification[96] serve as promising methods in the prediction of phishing attacks.

Phishing attack typically starts by, attacker sending an email to victims requesting personal information to be disclosed, by visiting a particular URL[97]. Phishers use a set of mutual features to create phishing websites to carry out proper deception[98]. We can exploit this information to successfully distinguish between phishy and non-phishy websites based on the extracted features of the website visited[94]. The two approaches that are commonly used in the identification of phishing sites are: black-list based, which involves comparison of the requested URL with those that are present in that list and Heuristic based method that involves the collection of certain features from the website to label it either as phishy or legitimate[99]. The disadvantage of Black-list based approach is that the black-list can not contain all phishing websites since, a new malicious website is launched every second[100]. In contrast, a Heuristic-based approach can recognize fraudulent websites that are new[101]. The success of Heuristic-based methods depend on the selection of features and the way they are processed. Data mining can be effectively used here to find patterns as well as relations among them[102]. Data mining is considered to be important for taking decisions, since decisions are made based on the patterns and rules derived using the data mining algorithms[103].

Although there is substantial progress made in the development of prevention techniques, phishing still remains a threat since most of the counter measures techniques in use are based still on reactive URL black-listing[104]. Since Phishing Web sites will have shorter life time these methods are considered to be inefficient. Newer approaches such as Associative Classification (AC) are more suitable for these kinds of applications. Associative Classification technique is a new technique derived by combining Association rule and Classification techniques of data mining[105]. AC typically includes two phases; the training phase to induce hidden knowledge (rules) using Association rule and the Classification phase to construct a Classifier after pruning useless and redundant rules. Many research studies have revealed that AC usually shows better classifiers with reference to error rate than other standard classification approaches such as decision tree and rule induction.

VII. Artificial Neural Networks(ANN)

An Artificial Neural Network is basically a connected set of processing units. Each connection has a specific weight that determines how one unit affects the other. Few of these units act as input nodes and few other as output nodes and remaining nodes consists of hidden layer. Neural network performs functionally, a mapping from input values to output values by activating each input node and allowing it to spread through the hidden layer nodes to the output nodes. The mapping is stored in terms of weight over connection. Fig. 7 shows the structure of HHNN[62].

ANN is one of the widely used techniques in the field of intrusion detection. ANN techniques are classified into three categories namely:

1. Supervised Intrusion Detection,
2. Unsupervised Intrusion Detection, and
3. Hybrid Intrusion detection.

Supervised Intrusion Detection based on ANN includes Multi Layer Feed Forward (MLFF) Neural Networks and Recurrent Neural Networks. Since,the number of training sets is huge and their distribution is imbalanced, the MLFF neural networks can easily reach the local minimum and hence have lower stability. The precision rate of a MLFF neural network is low for less frequent attacks. Supervised IDS exhibits lower detection performance than SVM and Multivariate Adaptive Regression Splines(MARS). Unsupervised Intrusion Detection based on ANN classifies test data and separates normal behaviors from abnormal ones. Since it does not need retraining, it can greatly improve the analysis of new data. The performance of Unsupervised ANN is also lower for low frequent attacks achieving a lower detection precision. Hybrid ways of combinining supervised ANN and unsupervised ANN and combining ANN with other data mining techniques for the detection of intrusion can be achieved to overcome the limitations of the basic types of ANN. A hybrid approach involving SOM and Radial Basis Function(RBF) networks is comparitively more efficient than Intrusion Detection based on RBF networks alone. A hybrid model that uses a combination of Flexible Neural Tree, Evolutionary Algorithm and Particle Swarm Optimization (PSO) is highly efficient. Hybrid ANN that uses a combination of Fuzzy Clustering technique with ANN reduces the training set into subsets that are smaller in size, thereby improving the stability of individual ANN for low-frequent attacks. So we can say that for intrusion detection based on ANNs, hybrid ANN has been the trend. Different ways of constructing Hybrid ANN influences the performance of intrusion detection. Hence it is required to construct different Hybrid ANN models to serve different goals. There are various Hybrid approaches being utilized for intrusion detections and one such model is the Hyperbolic Hopfield Neural Network(HHNN). Anomaly detection assumes that the intrusions always return as a number of deviations from the normal patterns. HHNN technique studies the relationship between the two sets of information, and generalizes it in getting new input-output pairs reasonably. Neural networks can be used hypothetically for the identification of attacks and look for these attacks in the audit stream. Since there is no reliable method at present to realize causes of association, it cannot clarify the reason behind the classification of the attack. The research progress made in HHNN is summarized in Table 3.

VIII. Anomaly Detection/Outlier Detection

Anomaly detection is a process that involves finding nonconforming patterns to the expected behavior. Such patterns are called *anomalies*. Different application domains term them differently as outliers or aberration or surprises or peculiarities or

Table 3: Research Progress in ANN

Authors	Algorithm	Performance	Future enhancement
C Cortes et al.(2016)[106]	Theoretical framework for analyzing and learning artificial neural networks	Optimizes generalization performance	Can be applied for different optimization tecniques and network architectures.
D T Bui et al.(2015)[107]	ROC and Kappa Index	MLP (90.2 %), SVM (88.7 %), KLR (87.9 %), RBF (87.1 %) and LMT (86.1 %).	Information Gain Ratio as feature selection can be tried.

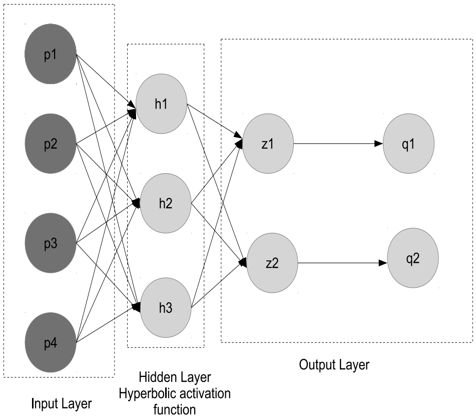


Figure 7: An overview of a HHNN

contaminants. Anomalies and Outliers are the two commonly used terms in this context. Anomaly detection applications are fraud detection of credit or debit cards, health care and insurance. It is also used for intrusion detection and fault detection in a safety critical system and for the detection of enemy activities.

Anomalous patterns mostly deviate from the normal patterns. The figure shown in Fig. 8 plots anomalies on a two dimensional data set. The regions *N1* and *N2* are considered normal, because majority of the observations lie in these regions. Points *O1* and *O2* that are far away from these regions and points in region *O3*, are considered anomalies. Although anomalies are induced in the data for a number of reasons, all of them have the common characteristic that they are interesting to analyze. This interestingness of outliers is a prime feature of anomaly or outlier detection. Anomaly detection is related to, but not as same as noise removal and noise accommodation, but it must ceratinly deal with unwanted noise in the data. Noise is an unwanted part to the analyst and acts as a hurdle to data analysis. Noise removal is therefore necessary, since unwanted data must be removed before performing the data analysis. Novelty Detection is a topic related to anomaly detection, which detects any previously unidentified novel patterns in the data. The detected novel patterns are incorporated into the normal model, that makes it different from Anomaly Detection. Different solutions that exist for anomaly detection will also work for novel Detection and *vice versa*. Hence in Anomaly Detection a region is defined, where the observations conforming to the region are considered normal and the non-conforming observations are considered anamolous.

a) Challenges

Some of the challenges the researchers face with respect to Anomaly Detection are:

1. Defining a normal region, where all normal behaviors exist is difficult. The boundary between normal and anomalous behavior has a very thin differentiation, meaning that an observation that lies closer to the boundary could be normal, and *vice versa*.
2. When the attackers masquerade to make the anomalous observations to appear normal, defining normal behavior becomes complicated.
3. Normal behaviors evolve and what is currently considered as normal might not be the same in the future.
4. Different application domains have different notions of anomaly. For instance fluctuations in body temperature marks an anomaly in the medical domain, while the fluctuations in marketing domain might be considered as normal. Therfore the application of a technique developed, cannot be generic.
5. Labeled data used by anomaly detection techniques for training/validation of models is not available freely. It is challenging to distinguish and remove noise from the data. The anomaly detection issue, is therefore hard to tackle with. Most of the anomaly detection techniques that exist, can only solve a problem formulation that is domain specific and is induced by factors such as category of the data, labeled data availability, anomaly types, and so on.

b) Data Mining Mechanisms for Anomaly Detection

An Intrusion Detection System can generally be implemented using the following two techniques:

1. Signature Based IDS and
2. Anomaly Based IDS.

Signature Based IDS makes use of Attack signatures that are explicitly defined and detect intrusions by Blacklisting.

It is ineffective against new types of attacks which makes it susceptible to evasion methods.

Anomaly Based IDS on the other hand, records normal behavior and classifies the deviations from normal behavior as anomalies. It is considered to be robust and reliable to unknown attacks and prevent attacks from malicious users who improvise their attacking strategy. The widely used implementation of Anomaly Based IDS is by the extensive use of data mining algorithms involving in two phases:

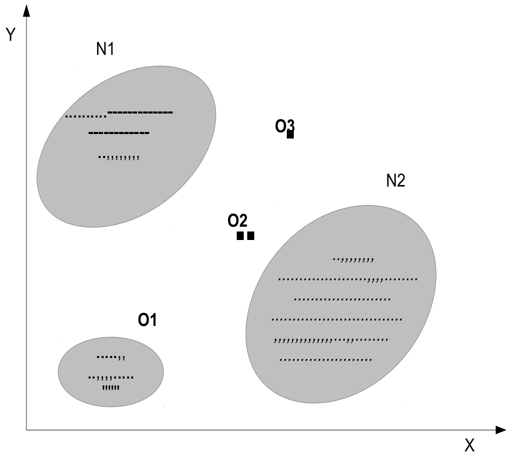


Figure 8: Outlier Detection

1. The Training Phase
2. The Detection Phase

During training phase, profiles are created by grouping normal access behaviors and are forwaded in a Batch mode to the Feature Extractor, Feature selector and Classifier. The Classifier produces a trained model out of normal access behavior[108]. Every new test sample during the Detection phase, is made to go through the same modules: Feature Extractor and Feature Selector, that is finally evaluated by the already trained Classifier. When the sample is found to be deviating from normal profiles, an alarm is raised. The profiles are required to be updated at regular intervals of time and Classifier training is also carried out periodically, so as to minimize the false alarm rate. For Feature selection, we can either employ the Ranking methods or the Filter methods. The Ranking methods output the feature set sorted in descending order according to a particular evaluation measure. The top variables in the feature set are considered to be the most discriminant features. It is therefore essential to determine a threshold to discard features that are considered to have little or no contribution to the classification process. Information Gain(IG) is one of the commonly used evaluation measures.

A variant of IG, with improvisation is the Gain Ratio (GR).

The GR overcomes the bias found in IG towards features resulting in a smaller set of features. For the purpose of Feature Selection we can employ a ranking method that is unsupervised called Principal components analysis(PCA).

The advantage of Filter methods for Feature Selection is that they automatically choose a set of selected features based on a particular evaluation measure. One of the widely employed Filtering methods for Feature Selection is the Best First Search(BFS). It makes use of Forward Selection and Backward Elimination to search through the feature space adopting a Greedy approach. When performance is found to be dropping, it backtracks to the previous feature subsets that have better performance and start all over again from there. BFS is computationally expensive for larger sets. Genetic Algorithms[109] is another type of Filtering technique that is considered to be very effective in practice[110].

IX. Mitigating Code Injection Attacks

A code injection attack typically involves writing of new machine code into the vulnerable programs memory[111], and after exploiting a bug in the program the control is redirected to the new code[112]. The protection technique[113], W+X mitigates this attack by allowing only either a Write or Execute operations on memory but never allows both[114].

The research progress made so far in this regard is summarized in Table 4.

a) *Types of Code Injection*

Some of the flavours of Code Injection attacks are: SQL Injection[121], HTML Script Injection[122], Object Injection[123], Remote File Injection[124] and Code Reuse Attacks(CRAs)[125].

i. *SQL Injecton*

A technique that uses SQL syntax to input commands that can alter read or modify a database is called SQL Injection. Consider for example a web page having a field on it to allow users to enter a password for authentication. The code behind the page usually a script code, will generate a SQL query to verify the matching password entered against the list of user names:

SELECT UsrList.Username FROM UsrList WHERE UsrList. Password = 'Password'

The access is granted when the password entered by the user matches the password specified in the query. If the malicious user can inject some valid code ('password' OR '1'='1') in the Password field. An attacker by leaving the password field empty makes the condition "'1'='1"' to become true and gains access to the database.

ii. *HTML Script Injection*

An attacker injects malicious code by making use of the <script>and </script>tags, within which he would change the location property of the document by setting it to an injected script.

iii. *Object Injection*

PHP allows serialization and deserialization of objects. If an untrustworthy input is allowed into the deserialization function, it is possible to modify existing classes in the program and execute malicious attacks.

iv. *Remote File Injection*

Attackers might provide a Remote Infected file name as the path by modifying the path command of the script file to cause the intended destruction[126].

Table 4: Research Progress in Code Injection Attacks

Authors	Algorithm	Performance	Future enhancement
M Graziano et al.(2016)[115]	Emulation-based framework for ROP	Total analysis time of 4 hours with 16-Core Intel E5-2630 (2.3GHz) and 24GB RAM	Reduction in total analysis time.
Mitropoulos et al.(2016)[116]	Contextual Fingerprinting	Overhead of 11.1% on execution time.	Overhead can be reduced.
A Follner et al.(2015)[117]	Dynamic Binary Instrumentation	2.4x overhead, comparable to similar approaches but no false alarms	The overhead can be reduced
G Parmar et al.(2015)[118]	Input based approach		More techniques/tools for SQLi prevention can be explored or created.
L Deng et al.(2015)[119]	Exception Oriented Programming (EOP)	Detection rate of about 90percent	Can be extended to Mac and Windows kernels.
S Gupta et al. (2015)[120]	Cross-Site Scripting Secure Web Application Framework	Ranging from 1.25% to 5.75% based on the type of JSP program	Discovering the techniques of dropping the HTTP response delay and other rule checks of XSS-SAFE without disturbing its efficiency of XSS attack recognition.

v. *Code Reuse Attacks*

Attacks in which an attacker directs control flow through an already existing code with an erroneous result are called Code Reuse Attacks[127].

Attackers therfore have come out with code-reuse attacks[128], in which a defect in the software is exploited to create a control flow through existing code-base to a malicious end[129]. The Return Into Lib C(RILC)is a type of code-reuse attack [130] where the stack is compromised and the control is transferred to the beginning of an existing library function such as *mprotect()* to create a memory region[131]that allows both write and execution operations on it to bypass W+X[132]. Such attacks can be effiently overcome using Data Mining techniques[133]. The source code is checked to find any such flaws and if so the instructions are classified as malicious[134]. Some of the classification Algorithms that can be used in this Regard are Bayesian[135], SVM[136] and Decision Tree[137].

vi. *Return Oriented Programming*

ROP attacks start when an attacker gains stack control[138] and redirects the control to a small snippet of code called gadget typically ending with a RET instruction[139]. Because attackers gain control over the return addresses[140], they can assign the RET of one gadget to the start of another gadget[141], achieving the desired functionality out of a large finite set of such small gadgets[142]. ROP Attacks inject no code and yet can induce arbitrary behavior in the targeted system [143]. A compiler-based approach has been suggested in [144] to combat any form of ROP. In [145], the authors present in-place code randomization that can be applied directly on third-party software, to mitigate ROP attacks. Buchanan et al., [146], have demonstrated that return-oriented exploits are practical to write, as the complexity of gadget combination is abstracted behind a programming language and compiler. Davi et al.[147] proposed runtime integrity monitoring techniques that use tracking instrumentation of program binaries based on taint analysis and dynamic tracing. In[148] a tool DROP, that detects ROP malicious code dynamically, is presented.

vii. *Jump Oriented Programming*

In Jump Oriented Programming(JOP), an attacker links the gadgets using a finite set of indirect JMP instructions[149], instead of RET instructions. A special gadget called a *dispatcher* is used for flow control management among the gadgets[150].

X. Conclusion

The purpose of this survey is to explore the importance of Data Mining techniques in achieving security. The paper is limited to few applications such as Privacy Preserving Data Mining (PPDM), Intrusion Detection System(IDS), Phishing Website Classification, Anomaly/Outlier Detection and Mitigation of Code Injection and Reuse Attacks. Some of the Classification and Clustering algorithms are discussed here considering their significance in Intrusion/Anomaly/ Outlier Detection Techniques. Other basic Data mining techniques such as Feature Extraction, Association Rule Mining and Decision Trees are also discussed, since many researchers have extensively used these techniques for IDS. The Survey could be made more exhaustive by exploring other security applications of Data Mining such as Malware Detection, Spam Detection, Web Mining and Crime Profiling.

References Références Referencias

1. D. R. Stinson, “Cryptography: Theory and Practice 3rd Edition,”*Text Book*, 2006.
2. C.-H. Yeh, G. Lee, and C.-Y. Lin, “Robust Laser Speckle Authentication System through Data Mining Techniques,” *IEEE Transactions on Industrial Informatics*, vol. 11, no. 2, pp. 505–512, 2015.
3. S. Khan, A. Sharma, A. S. Zamani, and A. Akhtar, “Data Mining for Security Purpose & its Solitude Suggestions,” *International Journal of Technology Enhancements and Emerging Engineering Research*, vol. 1, no. 7, pp. 1–4, 2012.

4. Venugopal K R, K G Srinivasa and L M Patnaik,"Soft Computing for Data Mining Applications," *Springer*, 2009.
5. Vasanthakumar G U, Bagul Prajakta , P Deepa Shenoy, Venugopal K R and L M Patnaik,"PIB: Profiling Influential Blogger in Online Social Networks, A Knowledge Driven Data Mining Approach,"*11th International Multi-Conference on Information Processing (IMCIP)*, vol. 54, pp. 362–370, 2015.
6. H. Zang and J. Bolot, "Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study," *In Proceedings of the 17th Annual International Conference on Mobile Computing and Networking*, pp. 145–156, 2011.
7. R. J. Bayardo and R. Agrawal, "Data Privacy Through Optimal *k*-Anonymization," *21st International Conference on Data Engineering (ICDE'05)*, pp. 217–228, 2005.
8. A. Friedman, R. Wolff, and A. Schuster, "Providing *k*-Anonymity in Data Mining," *The VLDB Journal*, vol. 17, no. 4, pp. 789–804, 2008.
9. R. Lu, X. Liang, X. Li, X. Lin, and X. Shen, "EPPA: An Efficient and Privacy-Preserving Aggregation Scheme for Secure Smart Grid Communications," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 9, pp. 1621–1631, 2012.
10. C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating Noise to Sensitivity in Private Data Analysis," *Theory of Cryptography Conference*, pp. 265–284, 2006.
11. M. Siddiqui, M. C. Wang, and J. Lee, "Detecting Internet Worms Using Data Mining Techniques," *Journal of Systemics, Cybernetics and Informatics*, vol. 6, no. 6, pp. 48–53, 2009.
12. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 Algorithms in Data Mining," *Knowle - dge and Information Systems*, vol. 14, no. 1, pp. 1–37, 2008.
13. M. S. Abadeh, J. Habibi, and C. Lucas, "Intrusion Detection using a Fuzzy Genetics-Based Learning Algorithm," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 414–428, 2007.
14. K. S. Desale and R. Ade, "Genetic Algorithm Based Feature Selection Approach for Effective Intrusion Detection System," *International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, 2015.
15. WU Bin, LU Tianliang, ZHENG Kangfeng, ZHENG Dongmei and LIN Xing,"Smartphone Malware Detection Model Based on Artificial Immune System,"*China Communications*, vol. 11, no. 13, pp. 86–92, 2014.
16. P Deepa Shenoy, Srinivasa K G, Venugopal K R and L M Patnaik, "Dynamic Association Rule Mining using Genetic Algorithms," *Intelligent Data Analysis*, vol. 9, no. 5, pp. 439–453, 2005.
17. P Deepa Shenoy, Srinivasa K G, Venugopal K R and L M Patnaik, "Evolutionary Approach for Mining Association Rules on Dynamic Databases,"*7th Pacific-Asia Conference on Knowledge Discovery and Data Mining(PAKDD)2003, Seoul, South Korea*, pp. 325–336, 2003.
18. S. J. Rizvi and J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," *Proceedings of the 28th International Conference on Very Large Data Bases*, pp. 682–693, 2002.
19. S. M. Darwish, M. M. Madbouly, and M. A. El-Hakeem, "A Database Sanitizing Algorithm for Hiding Sensitive Multi-level Association Rule mining," *International Journal of Computer and Communication Engineering*, vol. 3, no. 4, pp. 285–293, 2014.
20. J. Vaidya and C. Clifton, "Privacy Preserving Association Rule Mining in Vertically Partitioned Data," *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 639–644, 2002.
21. J. Vaidya and C. Clifton, "Secure Set Intersection Cardinality with Application to Association Rule Mining," *Journal of Computer Security*, vol. 13, no. 4, pp. 593–622, 2005.
22. M. R. B. Diwate and A. Sahu, "Efficient Data Mining in SAMS through Association Rule," *International Journal of Electronics Communication and Computer Engineering*, vol. 5, no. 3, pp. 593–597, 2014.
23. F. Thabtah, P. Cowling, and Y. Peng, "MCAR: Multi-class Classification based on Association Rule," *The 3rd ACS/IEEE International Conference on Computer Systems and Applications*, pp. 33–39, 2005.
24. K. Hu, Y. Lu, L. Zhou, and C. Shi, "Integrating Classification and Association Rule Mining: A Concept Lattice Framework," *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, pp. 443–447, 1999.
25. M. Hussein, A. El-Sisi, and N. Ismail, "Fast Cryptographic Privacy Preserving Association Rules Mining on Distributed Homogenous Database," *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 607–616, 2008.
26. J. Zhan, S. Matwin, and L. Chang, "Privacy-Preserving Collaborative Association Rule Mining," *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 153–165, 2005.
27. F. Giannotti, L. V. Lakshmanan, A. Monreale, D. Pedreschi, and H. Wang, "Privacy-Preserving Mining of Association Rules from Outsourced Transaction Databases," *IEEE Systems Journal*, vol. 7, no. 3, pp. 385–395, 2013.
28. K. Ilgun, R. A. Kemmerer, and P. A. Porras, "State Transition Analysis: A Rule-Based Intrusion Detection Approach," *IEEE Transactions on Software Engineering*, vol. 21, no. 3, pp. 181–199, 1995.

30. V. Kumar, H. Chauhan, and D. Panwar, "K-Means Clustering Approach to Analyze NSL-KDD Intrusion Detection Dataset," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 3, no. 4, pp. 1–4, 2013.
31. M. Taylor, "Data Mining with Semantic Features Represented as Vectors of Semantic Clusters," *Springer-Verlag*, pp. 1–16, 2012.
32. S. S. Shapiro, "Situating Anonymization within a Privacy Risk Model," *2012 IEEE International Systems Conference(SysCon)*, pp. 1–6, 2012.
33. A.C. Yao, "Protocols for Secure Computations," *23rd Annual Symposium on Foundations of Computer Science, 1982. SFCS'08*, pp. 160– 164, 1982.
34. A. Ben-David, N. Nisan, and B. Pinkas, "FairplayMP: A System for Secure Multi-Party Computation," *Proceedings of the 15th ACM Conference on Computer and Communications Security*, pp. 257–66, 2008.
35. P. Mohan, A. Thakurta, E. Shi, D. Song, and D. Culler, "GUPT: Privacy Preserving Data Analysis made Easy," *In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 349–360, 2012.
36. A. Kiayias, S. Xu, and M. Yung, "Privacy preserving Data Mining within Anonymous Credential Systems," *International Conference on Security and Cryptography for Networks*, pp. 57–76, 2008.
37. L. A. Dunning and R. Kresman, "Privacy Preserving Data Sharing with Anonymous ID Assignment," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 2, pp. 402–413, 2013.
38. M. Ouda, S. Salem, I. Ali, and E.-S. Saad, "Privacy-Preserving Data Mining in Homogeneous Collaborative Clustering," *International Arab Journal of Information Technology (IAJIT)*, vol. 12, no. 6, pp. 604– 612, 2015.
39. J. Vaidya and C. Clifton, "Privacy-Preserving Data Mining: Why, How, and When," *IEEE Security & Privacy*, vol. 2, no. 6, pp. 19–27, 2004.
40. B. Pinkas, "Cryptographic Techniques for Privacy-Preserving Data Mining," *ACM SIGKDD Explorations Newsletter*, vol. 4, no. 2, pp. 12–19, 2002.
41. M. Roughan and Y. Zhang, "Privacy-Preserving Performance Measurements," *In Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data*, pp. 329–334, 2006.
42. W. Du and Z. Zhan, "Using Randomized Response Techniques for Privacy-Preserving Data Mining," *In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 505–510, 2003.
43. M. Kantarcioglu, C. Clifton *et al.*, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 9, pp. 1026–1037, 2004.
44. J. Zhan, "Privacy-Preserving Collaborative Data Mining," *IEEE Computational Intelligence Magazine*, vol. 3, no. 2, pp. 31–41, 2008.
45. D. Frey, R. Guerraoui, A. Kermarrec, A. Rault, Ta¨ıani, Franc¸ois and J. Wang, "Hide and Share: Landmark-Based Similarity for Private KNN Computation," *In Proceedings of the 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, pp. 263–274, 2015.
46. T. Zhu, P. Xiong, G. Li and W. Zhou, "Correlated Differential Privacy: Hiding Information in Non-IID Data Set," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 2, pp. 229–242, 2015.
47. BK. Samanthula, Y. Elmehdwi and W. Jiang," K-Nearest Neighbor Classification over Semantically Secure Encrypted Relational Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 5, pp. 1261–73, 2015.
48. N P Nethravathi, Prashanth G Rao, P Deepa Shenoy, Venugopal K R and Indramma M, "CBTS: Correlation Based Transformation Strategy for Privacy Preserving Data Mining," *2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE), Dhaka, Bangladesh*, pp. 190–194, 2015.
49. N. Mohammed, D. Alhadidi, B. Fung and M. Debbabi,"Secure Two- Party Differentially Private Data Release for Vertically Partitioned Data," *IEEE Transactions on Dependable and Secure Computing*, vol. 11, no. 1, pp. 59–71, 2014.
50. J. Vaidya, B. Shafiq, W. Fan, D. Mehmood and D. Lorenzi," A Random Decision Tree Framework for Privacy-Preserving Data Mining," *IEEE Transactions on Dependable and Secure Computing*, vol. 11, no. 5, pp. 399–411, 2014.
51. YJ. Lee,"Privacy-preserving Data Mining for Personalized Marketing," *International Journal of Computer Communications and Networks (IJCCN)*, vol. 4, no. 1, pp. 1–7, 2014.
52. N. Zhang, M. Li, and W. Lou, "Distributed Data Mining with Differential Privacy," *IEEE International Conference on Communications*, pp. 1–5, 2011.
53. F. McSherry and I. Mironov, "Differentially Private Recommender Systems: Building Privacy Into the Net," *In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 627–636, 2009.
54. A. Friedman and A. Schuster, "Data Mining with Differential Privacy," *In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 493–502, 2010.
55. M. Roughan and Y. Zhang, "Secure Distributed Data Mining and Its Application to Large Scale Network Measurements," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 7–14, 2006.

56. N P Nethravathi, Vaibhav J Desai, P Deepa Shenoy, M Indiramma and Venugopal K R,"A Brief Survey on Privacy Preserving Data Mining Techniques," *Data Mining and Knowledge Engineering*, vol. 8, no. 9, pp. 267–273, 2016.
57. A. Narayanan and V. Shmatikov, "De-Anonymizing Social Networks," *IEEE Symposium on Security and Privacy*, pp. 173–187, 2009.
58. S. Chourse and V. Richhariya, "Survey Paper on Intrusion Detection Using Data Mining Techniques," *International Journal of Emerging Technology and Advanced Engineering, ISO*, vol. 4, no. 8, pp. 653–657, 2008.
59. S. Kumar and E. H. Spafford, "A Software Architecture to Support Misuse Intrusion Detection," *Computer Science Technical Report, Purdue University*, pp. 1–19, 1995.
60. J. Allen, A. Christie, W. Fithen, J. McHugh, and J. Pickel, "State of the Practice of Intrusion Detection Technologies," *Technical Report*, pp. 1–239, 2000.
61. R. Mitchell and R. Chen, "Adaptive Intrusion Detection of Malicious Unmanned Air Vehicles Using Behavior Rule Specifications," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 5, pp. 593–604, 2014.
62. J. Jabez and B. Muthukumar, "Intrusion Detection System: Time Probability Method and Hyperbolic Hopfield Neural Network," *Journal of Theoretical & Applied Information Technology*, vol. 67, no. 1, pp. 65– 77, 2014.
63. E. K. P G Reddy, M. Iaeng, V. Reddy, and Rajulu, "A Study of Intrusion Detection in Data Mining," *World Congress on Engineering (WCE)*, pp. 6–8, 2011.
64. W. Lee, S. J. Stolfo, and K. W. Mok, "A Data Mining Framework for Building Intrusion Detection Models," *Proceedings of the 1999 IEEE Symposium on Security and Privacy*, pp. 120–132, 1999.
65. S. K. Sahu, S. Sarangi, and S. K. Jena, "A Detail Analysis on Intrusion Detection Datasets," *IEEE International on Advance Computing Conference(IACC)*, pp. 1348–1353, 2014.
66. A. A. C´ardenas, R. Berthier, R. B. Bobba, J. H. Huh, J. G. Jetcheva, D. Grochocki, and W. H. Sanders, "A Framework for Evaluating Intrusion Detection Architectures in Advanced Metering Infrastructures," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 906–915, 2014.
67. Y. Al-Nashif, A. A. Kumar, S. Hariri, Y. Luo, F. Szidarovsky, and G. Qu, "Multi-Level Intrusion Detection System," *International Conference on Autonomic Computing ICAC'08*, pp. 131–140, 2008.
68. D. Beaver, S. Micali, and P. Rogaway, "The Round Complexity of Secure Protocols," *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, pp. 503–513, 1990.
69. W. Lu and I. Traore, "Detecting New Forms of Network Intrusion Using Genetic Programming," *Computational Intelligence*, vol. 20, no. 3, pp. 475–494, 2004.
70. T. F. Lunt, A. Tamaru, and F. Gillham, "A Real-Time Intrusion- Detection Expert System (IDES)," *Technical Report*, pp. 1–166, 1992.
71. D. Barbara, J. Couto, S. Jajodia, L. Popyack, and N. Wu, "ADAM: Detecting Intrusions by Data Mining," *Proceedings of the IEEE Workshop on Information Assurance and Security*, pp. 11–16, 2001.
72. M. Shetty and N. Shekokar, "Data Mining Techniques for Real Time Intrusion Detection Systems," *International Journal of Scientific & Engineering Research*, vol. 3, no. 4, pp. 1–7, 2012.
73. W. Lee, S. J. Stolfo, and K. W. Mok, "Adaptive Intrusion Detection: A Data Mining Approach," *Artificial Intelligence Review*, vol. 14, no. 6, pp. 533–567, 2000.
74. E. Bloedorn, A. D. Christiansen, W. Hill, C. Skorupka, L. M. Talbot, and J. Tivel, "Data Mining for Network Intrusion Detection: How to Get Started," *MITRE*, pp. 1–9, 2001.
75. R. Gopalakrishna and E. H. Spafford, "A Framework for Distributed Intrusion Detection using Interest Driven Cooperating Agents," *Technical Report*, pp. 1–24, 2001.
76. B. Mukherjee, L. T. Heberlein, and K. N. Levitt, "Network Intrusion Detection," *IEEE Network*, vol. 8, no. 3, pp. 26–41, 1994.
77. R. Heady, G. F. Luger, A. Maccabe, and M. Servilla, "The Architecture of a Network Level Intrusion Detection System," *Academic Work submitted to the University of New Mexico*.
78. D. Barbara, N. Wu, and S. Jajodia, "Detecting Novel Network Intrusions using Bayes Estimators." *SDM*, pp. 1–17, 2001.
79. M. Roesch *et al.*, "SNORT: Lightweight Intrusion Detection for Networks," *Proceedings of LISA '99: 13th Systems Administration Conference*, pp. 229–238, 1999.
80. R. Mitchell and R. Chen, "Effect of Intrusion Detection and Response on Reliability of Cyber Physical Systems," *IEEE Transactions on Reliability*, vol. 62, no. 1, pp. 199–210, 2013.
81. D.-Y. Yeung and Y. Ding, "Host-Based Intrusion Detection Using Dynamic and Static Behavioral Models," *Pattern Recognition*, vol. 36, no. 1, pp. 229–243, 2003.
82. W. Lee, S. J. Stolfo, and K. W. Mok, "Mining Audit Data to Build Intrusion Detection Models," *KDD-98 Proceedings*, pp. 66–72, 1998.
83. S. Tanachaiwiwat and K. Hwang, "Differential Packet Filtering Against DDos Flood Attacks," *ACM Conference on Computer and Communications Security (CCS)*, pp. 1–15, 2003.
84. C.-Y. Tseng, P. Balasubramanyam, C. Ko, R. Limprasittiporn, J. Rowe, and K. Levitt, "A Specification-Based Intrusion Detection System

forAODV," *Proceedings of the 1st ACM Workshop on Security of Ad-hoc and Sensor Networks*, pp. 125–134, 2003.

85. M. S. Vittapu, V. Sunkari, and A. Y. Abate, "The Practical Data Mining Model for Efficient IDS Through Relational Databases," *International Journal of Research in Engineering and Science*, vol. 3, no. 1, pp. 20–30, 2015.
86. P. Soni and P. Sharma, "An Intrusion Detection System Based on KDD-99 Data Using Data Mining Techniques and Feature Selection," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 4, pp. 112–118, 2014.
87. M. Dubiner, Z. Galil, and E. Magen, "Faster Tree Pattern Matching," *Journal of the ACM (JACM)*, vol. 41, no. 2, pp. 205–213, 1994.
88. C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion Detection by Machine Learning: A Review," *Expert Systems with Applications*, vol. 36, no. 10, pp. 11 994–12 000, 2009.
89. D. M. Farid, N. Harbi, and M. Z. Rahman, "Combining Naive Bayes and Decision Tree for Adaptive Intrusion Detection," *arXiv preprint arXiv:1005.4496*, 2010.
90. A. J. Smola and B. Sch¨olkopf, "A Tutorial on Support Vector Regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.
91. T. Joachims, "Text Categorization with Support Vector Machines: Learning With Many Relevant Features," *European Conference on Machine Learning*, pp. 137–142, 1998.
92. X. Xu and X. Wang, "An Adaptive Network Intrusion Detection Method Based on PCA and Support Vector Machines," *International Conference on Advanced Data Mining and Applications*, pp. 696–703, 2005.
93. Vasanthakumar G U, P Deepa Shenoy, Venugopal K R and L M Patnaik,"PFU: Profiling Forum Users in Online Social Networks, A Knowledge Driven Data Mining Approach,"*2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECONECE)*, pp. 57–60, 2015.
94. A. Herzberg and A. Gbara, "Trustbar: Protecting (even naive) Web Users from Spoofing and Phishing Attacks," *Cryptology ePrint Archive Report 2004/155. http://eprint.iacr.org/2004/155*, 2004.
95. R. M. Mohammad, F. Thabtah, and L. McCluskey, "Intelligent Rule- Based Phishing Websites Classification," *IET Information Security*, vol. 8, no. 3, pp. 153–160, 2014.
96. S. Marchal, J. Franc¸ois, R. State, and T. Engel, "PhishStorm:Detecting Phishing with Streaming Analytics," *IEEE Transactions on Network and Service Management*, vol. 11, no. 4, pp. 458–471, 2014.
97. N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing Detection Based Associative Classification Data Mining," *Expert Systems with Applications*, vol. 41, no. 13, pp. 5948–5959, 2014.
98. Y. Zhang, S. Egelman, L. Cranor, and J. Hong, "Phinding Phish: Evaluating Anti-Phishing Tools," *Academic Work Submitted to School of Computer Science at Research Showcase @ CMU*.
99. M. Aburrous, M. A. Hossain, K. Dahal, and F. Thabtah, "Predicting Phishing Websites using Classification Mining Techniques with experimental Case Studies," *Seventh International Conference on Information Technology: New Generations (ITNG)*, pp. 176–181, 2010.
100. J. Chen and C. Guo, "Online Detection and Prevention of Phishing Attacks," *First International Conference on Communications and Networking in China*, pp. 1–7, 2006.
101. A. Y. Fu, L. Wenyin, and X. Deng, "Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (emd)," *IEEE Transactions on Dependable and Secure Computing*, vol. 3, no. 4, pp. 301–311, 2006.
102. T. Moore and R. Clayton, "An Empirical Analysis of the Current State of Phishing Attack and Defence," *Academic work*, 2007.
103. J. Hong, "The State of Phishing Attacks," *Communications of the ACM*, vol. 55, no. 1, pp. 74–81, 2012.
104. Asha S Manek, P Deepa Shenoy, M Chandra Mohan and Venugopal K R,"Detection of Fraudulent and Malicious Websites by Analysing User Reviews for Online Shopping Websites," *International Journal of Knowledge and Web Intelligence*, vol. 5, no. 3, pp. 171–189, 2016.
105. C. Jackson, D. R. Simon, D. S. Tan, and A. Barth, "An Evaluation of Extended Validation and Picture-in-Picture Phishing attacks," *International Conference on Financial Cryptography and Data Security*, pp. 281–293, 2007.
106. C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, S. Yang, "AdaNet: Adaptive Structural Learning of Artificial Neural Networks," *arXiv:1607.01097v1*, vol. 1, no. 17, pp. 1–18, 2016.
107. D. Bui, T. Tuan, H. Klempe, B. Pradhan, I. Revhaug, "Spatial Prediction Models for Shallow Landslide Hazards: A Comparative Assessment of the Efficacy of Support Vector Machines, Artificial Neural Networks, Kernel Logistic Regression, and Logistic Model Tree," *springer-Verlag Berlin Heidelberg*, vol. 13, no. 2, pp. 361–378, 2015.
108. M. Qin and K. Hwang, "Effectively Generating Frequent Episode Rules for Anomaly-based Intrusion Detection," *IEEE Symposium on Security and Privacy*, 2003.
109. Rasheed and R. Alhajj, "A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences," *IEEE Transactions on Cybernetics*, vol. 44, no. 5, pp. 569–582, 2014.

110. K. Hwang, P. Dave, and S. Tanachaiwiwat, "Netshield: Protocol Anomaly Detection with Datamining against DDOS Attacks," *Proceedings of the 6th International Symposium on Recent Advances in Intrusion Detection, Pittsburgh, PA*, pp. 8–10, 2003.
111. X. Liu, P. Zhu, Y. Zhang, and K. Chen, "A Collaborative Intrusion Detection Mechanism Against False Data Injection Attack in Advanced Metering Infrastructure," *IEEE Transactions on Smart Grid*, vol. 6, no. 5, pp. 2435–2443, 2015.
112. M. G. Schultz, E. Eskin, F. Zadok, and S. J. Stolfo, "Data Mining Methods for Detection of New Malicious Executables," *In Proceedings of IEEE Symposium on Security and Privacy S&P*, pp. 38–49, 2001.
113. B. Wu, T. Lu, K. Zheng, D. Zhang, and X. Lin, "Smartphone Malware Detection Model Based on Artificial Immune System," *China Communications*, vol. 11, no. 13, pp. 86–92, 2015.
114. D. K. B. Patel and S. H. Bhatt, "Implementnig Data Mining for Detection of Malware from Code," *An International Journal of Advanced Computer Technology:Compusoft*, vol. 3, no. 4, pp. 732–740, 2014.
115. Mariano Graziano, Davide Balzarotti and Alain Zidouemba, "ROPMEMU: A Framework for the Analysis of Complex Code-Reuse Attacks," *11th ACM Asia Conference on Computer and Communications Security*, pp. 1–12, 2016.
116. D Mitropoulos, K Stroggylos and D Spinellis, "How to Train Your Browser: Preventing XSS Attacks Using Contextual Script Fingerprints," *ACM Transactions on Privacy and Security*, vol. 19, no. 1, pp. 1–31, 2016.
117. Follner, E. Bodden, "ROPocop - Dynamic Mitigation of Code- Reuse Attacks," *Secure Software Engineering Group*, vol. 29, no. 3, pp. 16–26, 2015.
118. G Parmar and Dr. Kirti Mathur, "Proposed Preventive measures and Strategies Against SQL injection Attacks," *Indian Journal of Applied Research*, vol. 5, no. 5, pp. 716–718, 2015.
119. L. Deng, Q. Zeng, "Exception-Oriented Programming: Retrofitting Code-Reuse Attacks to Construct Kernel Malware," *The Institution of Engineering and Technology*, vol. 10, no. 6, pp. 418–424, 2016.
120. S. Gupta B.B. Gupta, "XSS-SAFE:A Server-Side Approach to Detect and Mitigate Cross-Site Scripting (XSS) Attacks in JavaScript Code," *Springer*, vol. 4, no. 3, pp. 897–920, 2015.
121. M. Polychronakis, "Generic Detection of Code Injection Attacks using Network-Level Emulation," *Ph.D. Thesis*, 2009.
122. P. Rauzy and S. Guilley, "A Formal Proof of Countermeasures Against Fault Injection Attacks on CRT-RSA," *Journal of Cryptographic Engineering*, vol. 4, no. 3, pp. 173–185, 2014.
123. S. Bhatkar, D. C. DuVarney, and R. Sekar, "Address Obfuscation: An Efficient Approach to Combat a Broad Range of Memory Error Exploits," *Usenix Security*, vol. 3, pp. 105–120, 2003.
124. E. G. Barrantes, D. H. Ackley, T. S. Palmer, D. Stefanovic, and D. D. Zovi, "Randomized Instruction Set Emulation to Disrupt Binary Code Injection Attacks," *Proceedings of the 10th ACM Conference on Computer and Communications Security*, pp. 281–289, 2003.
125. S. Bhatkar, D. C. DuVarney, and R. Sekar, "Efficient Techniques for Comprehensive Protection from Memory Error Exploits," *Proceedings of the 14th USENIX Security Symposium*, 2005.
126. Venugopal K R and Rajkumar Buyya,"Mastering C++," *Tata McGraw- Hill Education*, 2013.
127. J. Habibi, A. Panicker, A. Gupta, and E. Bertino, "DISARM: Mitigating Buffer Overflow Attacks on Embedded Devices," *International Conference on Network and System Security*, pp. 112–129, 2015.
128. M. Kayaalp, T. Schmitt, J. Nomani, D. Ponomarev, and N. A. Ghazaleh, "Signature-Based Protection from Code Reuse Attacks," *IEEE Transactions on Computers*, vol. 64, no. 2, pp. 533–546, 2015.
129. E. G¨oktas¸, E. Athanasopoulos, M. Polychronakis, H. Bos, and G. Portokalidis, "Size Does Matter:Why Using Gadget-Chain Length to Prevent Code-Reuse Attacks is Hard," *USENIX Security Symposium*, pp. 417–432, 2014.
130. S. Kevin Z, F. Monrose, D. Fabian, D. Lucas, L. Alexandra, S. Christopher, and R. Ahmad, "Just-In-Time Code Reuse: On the Effectiveness of Fine-Grained Address Space L]ayout Randomization, " *2013 IEEE Symposium on Security and Privacy*, pp. 574–588, 2013.
131. V. van der Veen, E. G¨oktas, M. Contag, A. Pawlowski, X. Chen, S. Rawat, H. Bos, T. Holz, E. Athanasopoulos, and C. Giuffrida, "A Tough Call: Mitigating Advanced Code-Reuse Attacks at the Binary Level," *IEEE Symposium on Security and Privacy*, pp. 1–20, 2016.
132. E. R. Jacobson, A. R. Bernat, W. R. Williams, and B. P. Miller, "Detecting Code Reuse Attacks with a Model of Conformant Program Execution," *International Symposium on Engineering Secure Software and Systems*, pp. 1–18, 2014.
133. M. Musuvathi, D. Y. Park, A. Chou, D. R. Engler, and D. L. Dill, "CMC: A Pragmatic Approach to Model Checking Real Code," *ACM SIGOPS Operating Systems Review*, vol. 36, no. 5, pp. 75–88, 2002.
134. N. Mohanappriya and R. Rajagopal, "Prediction and Pan Code Reuse Attack by Code Randomization Mechanism and Data Corruption," *Techniques and Algorithms in Emerging Technologies*, pp. 162–168, 2016.
135. D. M. Stanley, "CERIAS Tech Report 2013-19 Improved Kernel Security through Code Validation,

Diversification, and Minimization," *Ph.D. Thesis*, 2013.

136. Y. Zhuang, T. Zheng, and Z. Lin, "Runtime Code Reuse Attacks: A Dynamic Framework Bypassing Fine-Grained Address Space Layout Randomization," *SEKE*, pp. 609–614, 2014.
137. G. F. Roglia, L. Martignoni, R. Paleari, and D. Bruschi, "Surgically Returning to Randomized Lib (C)," *Computer Security Applications Conference*, pp. 60–69, 2009.
138. E. Buchanan, R. Roemer, H. Shacham, and S. Savage, "When Good Instructions Go Bad: Generalizing Return-Oriented Programming to RISC," *Proceedings of the 15th ACM Conference on Computer and Communi-cations Security*, pp. 27–38, 2008.
139. A. Gupta, S. Kerr, M. S. Kirkpatrick, and E. Bertino, "MARLIN: A Fine Grained Randomization Approach to Defend Against ROP Attacks," *International Conference on Network and System Security*, pp. 293–306, 2013.
140. S. Checkoway, L. Davi, A. Dmitrienko, A.-R. Sadeghi, H. Shacham, and M. Winandy, "Return-Oriented Programming Without Returns," *Proceedings of the 17th ACM Conference on Computer and Communications Security*, pp. 559–572, 2010.
141. L. Davi, A.-R. Sadeghi, and M. Winandy, "ROP Defender: A Detection Tool to Defend Against Return-Oriented Programming Attacks," *Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security*, pp. 40–51, 2011.
142. L. Davi, A. Dmitrienko, A.-R. Sadeghi, and M. Winandy, "Return- Oriented Programming Without Returns on ARM," *Technical Report HGI-TR-2010-002*, 2010.
143. R. Roemer, E. Buchanan, H. Shacham, and S. Savage, "Return- Oriented Programming: Systems, languages, and applications," *ACM Transactions on Information and System Security (TISSEC)*, vol. 15, no. 1, pp. 1–34, 2012.
144. K. Onarlioglu, L. Bilge, A. Lanzi, D. Balzarotti, and E. Kirda, "G-Free: Defeating Return-Oriented Programming Through Gadget- Less Binaries," *Proceedings of the 26th Annual Computer Security Applications Conference*, pp. 49–58, 2010.
145. V. Pappas, M. Polychronakis, and A. D. Keromytis, "Smashing the Gadgets: Hindering Return-Oriented programming using In-Place Code Randomization," *2012 IEEE Symposium on Security and Privacy*, pp. 601–615, 2012.
146. E. Buchanan, R. Roemer, H. Shacham, and S. Savage, "When Good Instructions Go Bad: Generalizing Return-Oriented Programming to RISC," *Proceedings of the 15th ACM Conference on Computer and Communications Security*, pp. 27–38, 2008.
147. L. Davi, A.-R. Sadeghi, and M. Winandy, "Dynamic Integrity Measurement and Attestation: Towards Defense Against Return-Oriented Programming Attacks," *Proceedings of the 2009 ACM Workshop on Scalable Trusted Computing*, pp. 49–54, 2009.
148. P. Chen, H. Xiao, X. Shen, X. Yin, B. Mao, and L. Xie, "Drop:Detecting Return-Oriented Programming Malicious Code," *International Conference on Information Systems Security*, pp. 163–177, 2009.
149. F. Yao, J. Chen, and G. Venkataramani, "JOP Alarm:Detecting Jump Oriented Programming Based Anomalies in Applications," *International Conference on Computer Design*, pp. 467–470, 2013.
150. T. Bletsch, X. Jiang, V. W. Freeh, and Z. Liang, "Jump-Oriented Programming: A New Class of Code-Reuse Attack," *Proceedings ACM Symposium on Information, Computer and Communications Security*, pp. 30–40, 2011.
151. S. Abadeh, A. Fernandez, A. Bawakid, S. Alshomrani and F. Herrera, "On The Combination of Genetic Fuzzy Systems and Pairwise Learning for Improving Detection Rates on Intrusion Detection Systems," *Journal of Expert Systems with Applications*, vol. 42, no. 1, pp. 193–202, 2015.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Estimation of Missing Attribute Value in Time Series Database in Data Mining

By Swati Jain & Mrs. Kalpana Jain

College of technology and Engineering,India

Abstract- Missing data is a widely recognized problem affecting large database in data mining. The substitution of mean values for missing data is commonly suggested and used in many statistical software packages, however, mean substitution lead to large errors in correlation matrix and therefore degrading the performance of statistical modeling. The problems arises are biasness of result data base, inefficient data in missing data when anomalous data is also present. In proposed work there is proper handling of missing data values and their analysis with removal of the anomalous data.This method provides more accurate and efficient result and reduces biasness of result for filling in missing data. Theoretical analysis and experimental results shows that proposed methodology is better.

Keywords : *missing data method, imputation, outlier, inference, anova test.*

GJCST-C Classification : *H.2.8,J.4*

Estimation of Missing Attribute Value in Time Series Database in Data Mining

Swati Jain [α] & Mrs. Kalpana Jain [σ]

Abstract- Missing data is a widely recognized problem affecting large database in data mining. The substitution of mean values for missing data is commonly suggested and used in many statistical software packages, however, mean substitution lead to large errors in correlation matrix and therefore degrading the performance of statistical modeling. The problems arises are biasness of result data base, inefficient data in missing data when anomalous data is also present. In proposed work there is proper handling of missing data values and their analysis with removal of the anomalous data.This method provides more accurate and efficient result and reduces biasness of result for filling in missing data. Theoretical analysis and experimental results shows that proposed methodology is better.

Keywords: missing data method, imputation, outlier, inference, anova test.

I. Introduction

Data cleaning is a step for discovery of database. Data cleaning, it is also known as data cleansing, it is a phase in which noisy data, anomalous data and irrelevant data are removed from the collection of various data. Missing data are defined as some of the values in the data set which are either lost or not observed or not available due to natural or non natural reasons. Data with missing values confuses both the data analysis and the submission of a solution to fresh data. Thus, three main problems arise when dealing with incomplete data. First, there is a loss of information and, as a consequence, a loss of efficiency. Second, there are several complications related to data handling, computation and analysis, due to the irregulaties in data structure and the impossibility of using standard software. Third, and most important, there may be bias due to systematic differences between observed and unobserved data. Deal with missing data is major task for cleaning data. Noor et all [1] In this paper, three types of mean imputation techniques introduced on missing data. Rubin [7] explored about inference and missing data and multiple imputations for non-response in the survey. Allison [8] investigated estimates of linear models with incomplete data and on missing data. Smyth [9] and Zhang [10] have considered that data preparation is a fundamental stage of data analysis. Therefore, this research focuses on anomalous and missing data values. In our research we create a novel method to replace the missing values.

II. Missing Data Methods

There are several methods for treating missing data. Missing data treatment methods can be divided into three categories, as proposed in [7].

a) *Ignoring and discarding data*

In this method the two main ways to discard data with missing values. The first method is known as list wise deletion. It consists of discarding all instances with missing data. The second method is known as pair wise deletion method. It consists of discarding instances or attributes before deleting any attribute, it is necessary to evaluate its relevance to the analysis.

b) *Parameter estimation*

In this missing data treatment method, Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm can handle parameter estimation in the presence of missing data.

c) *Imputation*

Imputation method is a class of procedures that aims to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set assist in estimating the missing values [9].

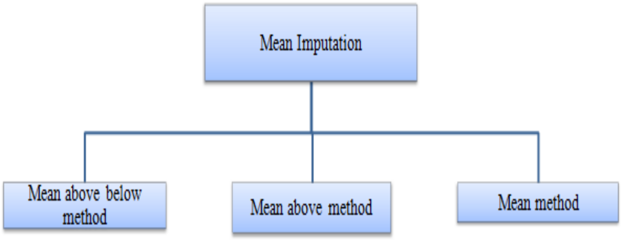


Figure 1: Types of mean imputation method

Mean Imputation Method: In this technique, It consists of replacing the missing data for a given feature by the mean of all known values of that attribute in the class where the instance with missing attribute belongs mean of each attribute that contains missing values is calculated and is replaced in the place of missing values. Each missing value is substituted with calculated mean value which is same for all.

Author α σ: College of technology and Engineering,India.
e-mail: swati.subhi.9@gmail.com

Mean Above Below Method: [1] this method replaces all missing values with the mean of the data above the missing value and one data below the missing value.

Mean Above Method [1]: This method replaces all missing values with the mean of all available data above the missing values.

Mean Method[1]: This method replaces all missing values with the mean of all available data.

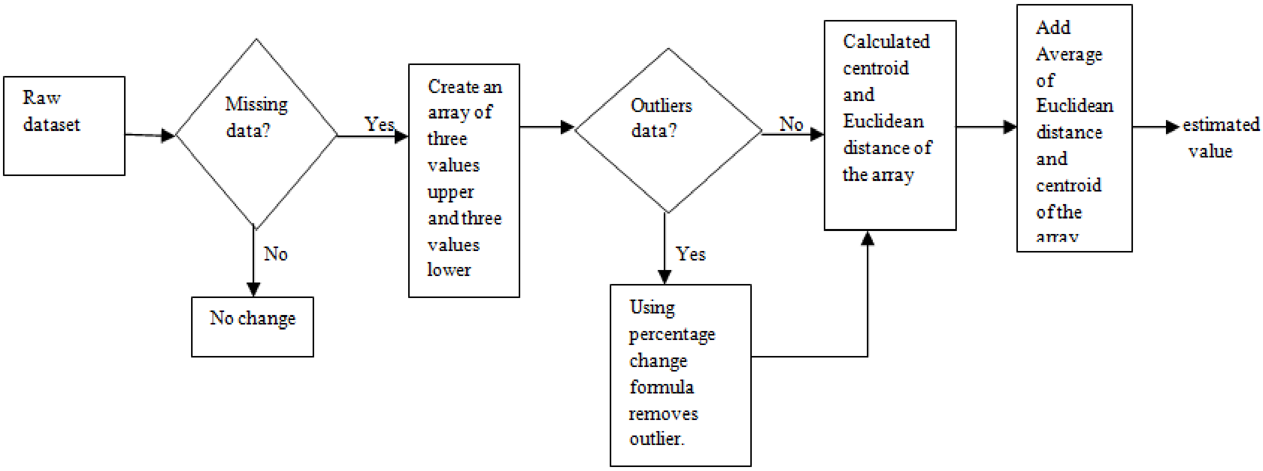

Raw dataset
Missing data?
Yes
No
No change
Create an array of three values upper and three values lower
Outliers data?
No
Yes
Using percentage change formula removes outlier.
Calculated centroid and Euclidean distance of the array
Add Average of Euclidean distance and centroid of the array
estimated value

Figure 2: Design flow of proposed methodology

III. Proposed Method for Inference of Missing Attributes Value in Data Mining

The proposed Methodology works in two stages. The first stage is localizing missing data and remove anomalous data, in next stage we substitute the estimated value in the place of missing values by using proposed method. This calculation gives the effective result and decreases the biasness of result.

The working stream of proposed work is shown in figure 2, if there are missing values in the raw data set, then the small subset/array is created from the input data sheet in which missing data value is existing, along with this we work out for anomalous value, according if anomalous value is presented, replace anomalous value with the new calculated value, last step of the work is estimation of missing values using Euclidean distance.

a) Missing value check & outlier detection

This step is preprocessed step of missing data in the input data sheet. The missing values locations are checked in entire data set.

As per the figure 3, the missing value case is pointed by the subscript of the attribute and denoted by the variable x_i. after pointing missing value case, we have to record the three upper value(x_{i-1},x_{i-2},x_{i-3}) and three lower value(x_{i+1},x_{i+2},x_{i+3}) from the missing value subscripts. Now the anomalous value in this subset is detected by the percentage change formula. After computing the percentage change of the subset. Now, we find the outlier range, value of outlier range define as per the suitable of the array value. If the anomalous value is detected in the data set, remove that value from the array.

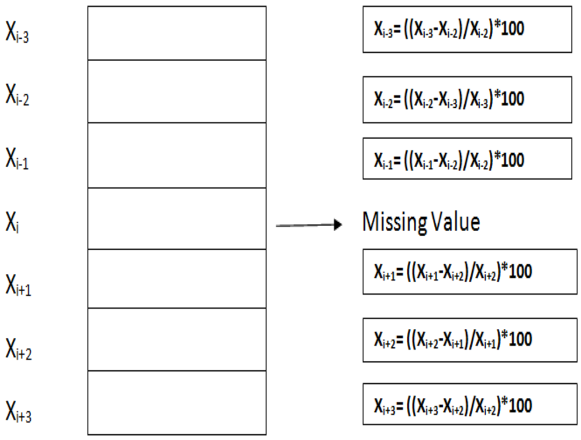

X_{i-3}
X_{i-2}
X_{i-1}
X_i
X_{i+1}
X_{i+2}
X_{i+3}
X_{i-3}= ((X_{i-3}-X_{i-2})/X_{i-2})*100
X_{i-2}= ((X_{i-2}-X_{i-3})/X_{i-3})*100
X_{i-1}= ((X_{i-1}-X_{i-2})/X_{i-2})*100
Missing Value
X_{i+1}= ((X_{i+1}-X_{i+2})/X_{i+2})*100
X_{i+2}= ((X_{i+2}-X_{i+1})/X_{i+1})*100
X_{i+3}= ((X_{i+3}-X_{i+2})/X_{i+2})*100

Figure 3: Calculation of percentage change for outlier detection

b) Calculation for missing values

Estimation of missing values in the last phase, when we have the outlier-free data. we process to fill missing values in this array, firstly calculate centroid of the subset ,centroid is generated by the mean of subset. At the further stage Euclidean distance is calculated between centroid of the data and the each value of the data. Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. The distance between vectors X and Y is defined as follows:

$$\text{Euclidean distance (d)} = \sqrt{\sum_{i=1}^{n}}(\text{Xa} - Ya)\,2$$

Here,
X_a is centroid of the array
Y_a is particular value of the array

at the last we compute the average of the Euclidean distance and add centroid with average value of distance this is the estimated value of the missing value. The value of X_{est}(estimated value) is separately computed for every missing value in the complete datasheet.

IV. Results and Discussion

Our experiments were carried out for time series datasets taken from Earthpolicy.org site. In proposed work we used different Datasheet like Hydroelectric Generation in India 1965-2013, Average Global Temperature 1880-2014, U.S. Motor Gasoline Consumption 1950-2014, World Wood Production 1961-2011 and few more. Here, we evaluate the U.S. Motor Gasoline Consumption 1950-2014 contains 50 number of instances and two attributes.

Below graph figure 4 shows comparison with respect to mean of all method. The U.S. Motor Gasoline Consumption respectively for the years 1950-2014 for million barrels attribute. The mean consumption of u.s. motor gasoline of million barrels are 2714.the variables are observed and missing values it may be noted that in the planned way 20% values are missing in the random manner for all the variables and in this dataset value of outliers is greater than 5.The mean calculated from incomplete data sets are 2379 this value is slightly lower than the mean values. The proposed methodology applied on the data sets to fill up missing values and the value is 2714.It is observed that the mean values are obtained after replacing missing values by proposed work are close to the actual mean. The results from the proposed method are compares with the techniques like mean above below(MAB), mean above(MA),mean imputation(MI), mean comparison method proposed by Noor et all [1] and analyze shows that proposed method value substitute missing values are more close to the original method with respect to the other method.

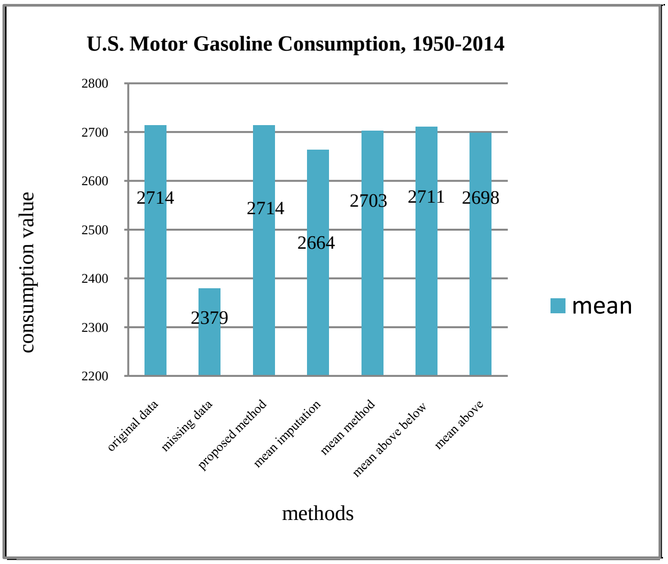


Figure 4: mean value

In figure 5 & figure 6 shows the comparison with respect to standard deviation value and coefficient of variance value of all methods. The proposed method performed significantly better than all other methods.

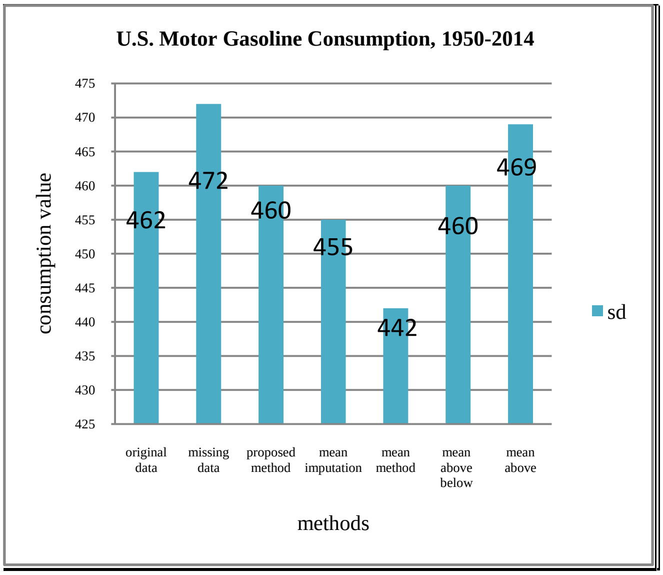


Figure 5: standard deviation value

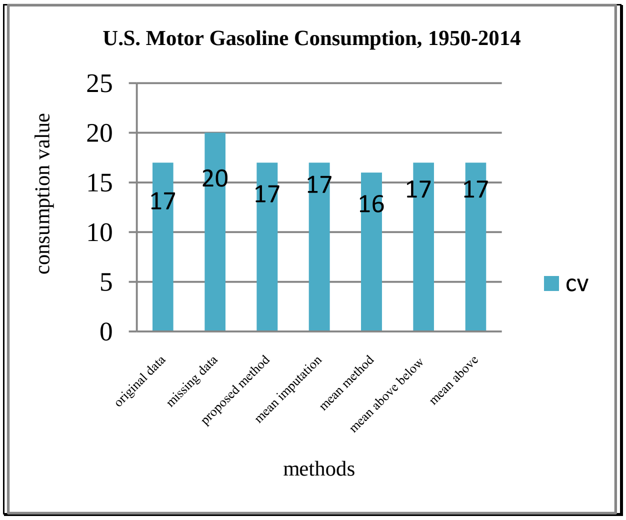


Figure 6: coefficient of variance value

Figure 4: Performance comparison of proposed methodology with MI (mean imputation), mean method, mean above below(MAB) method, mean above(MA) method. (a) Mean value (b) standard deviation value(c) coefficient of variance value

The data sheets are imported in the SPSS and necessary tests for the data validation and significance were applied. On the SPSS software the results are checked by using the ANOVA test for the data sheet and significance value is 99.1% that shows the result is efficient and more compatible with original data.

V. Conclusion

The work focuses on imputing missing values using proposed methodology for numerical attribute in time series data sheet. This method is suitable to handling missing data alone in presence of anomalous data. In this work, performance of proposed method is more reliable as comparing to other mean imputation technique for data analysis in the data mining field.

References Références Referencias

1. Noor, M. N., Yahaya, A. S., Ramli, N. A., & Al Bakri, A. M. M. 2014. Mean imputation techniques for filling the missing observations in air pollution dataset. *Key Engineering Materials* **594-599**: 902-908 Trans Tech Publications.
2. Cho, H. Y., Oh, J. H., Kim, K. O., & Shim, J. S. 2013. Outlier detection and missing data filling methods for coastal water temperature data. *Journal of Coastal Research*, **65**:1898-1903.
3. Porter, J. R., Cossman, R. E., & James, W. L. 2009. imputing large group averages for missing data,
4. using rural-urban continuum codes for density driven industry sectors. *Journal of Population Research*, ***26*****(3)**: 273-278.
5. Graham, J. W. 2009. Missing data analysis: Making it work in the real world. *Annual review of psychology*, **60**: 549-576.
6. Barnett, V., & Lewis, T. 1994. Outliers in Statistical Data ,John Wiley and Sons. *New York*.
7. Genolini, C., & Jacqmin-Gadda, H. 2013. Copy mean: a new method to impute intermittent missing values in longitudinal studies. *Open Journal of Statistics*, **3**: 26-40.
8. Rubin, D. B. 1976. Inference and missing data. *Biometrika*, ***63*****(3)**: 581-592.
9. Allison, P. D. 1987. Estimation of linear models with incomplete data. *Sociological methodology*, 71-103.
10. Smyth, P. 2001. Data mining at the interface of computer science and statistics. In *Data mining for scientific and engineering applications* 35-61. Springer US.
11. Zhang, S., Zhang, C., & Yang, Q. 2003. Data prepara tion for data mining. *Applied Artificial Intelligence*, **17(5-6):** 375-381.
12. Chen, L., Toma-Drane, M., Valois, R. F., & Drane, J. W. 2005. Multiple imputation for missing ordinal data. *Journal of Modern Applied Statistical Methods*, **4(1):** 26.
13. Zhu, X., Zhang, S., Jin, Z., Zhang, Z., & Xu, Z. 2011. Missing value estimation for mixed-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering*, ***23*****(1)**: 110-121.
14. Song, Q., & Shepperd, M. 2007. A new imputation method for small software project data sets. *Journal of Systems and Software*, **80(1)**: 51-62.
15. Grzymala-Busse, J. W. 2004. Data with missing attribute values: Generalization of indiscernibility relation and rule induction. *Transactions on Rough Sets* **1**: 78- 95. Springer Berlin Heidelberg.
16. Han, J., Pei, J., & Kamber, M. 2011. Data mining: concepts and techniques. Morgan Kaufmann Publishers, 225 Wyman Street,Waltham, USA pp. 83-91.

Global Journal of Computer Science and Technology: C
Software & Data Engineering
Volume 16 Issue 5 Version 1.0 Year 2016
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Sentiment Analysis and Opinion Mining from Social Media : A Review

By Savitha Mathapati, S H Manjula & Venugopal K R

University Visvesvaraya College of Engineering

Abstract - Ubiquitous presence of internet, advent of web 2.0 has made social media tools like blogs, Facebook, Twitter very popular and effective. People interact with each other, share their ideas, opinions, interests and personal information. These user comments are used for finding the sentiments and also add financial, commercial and social values. However, due to the enormous amount of user generated data, it is an expensive process to analyze the data manually. Increase in activity of opinion mining and sentiment analysis, challenges are getting added every day. There is a need for automated analysis techniques to extract sentiments and opinions conveyed in the user-comments. Sentiment analysis, also known as opinion mining is the computational study of sentiments and opinions conveyed in natural language for the purpose of decision making.

Keywords: domain adaptation; machine learning; opining mining; sentiment analysis; sentiment classification.

GJCST-C Classification : H.2.8,J.4

Sentiment Analysis and Opinion Mining from Social Media: A Review

Savitha Mathapati [α], S H Manjula [σ] & Venugopal K R [ρ]

Abstract- Ubiquitous presence of internet, advent of web 2.0 has made social media tools like blogs, Facebook, Twitter very popular and effective. People interact with each other, share their ideas, opinions, interests and personal information. These user comments are used for finding the sentiments and also add financial, commercial and social values. However, due to the enormous amount of user generated data, it is an expensive process to analyze the data manually. Increase in activity of opinion mining and sentiment analysis, challenges are getting added every day. There is a need for automated analysis techniques to extract sentiments and opinions conveyed in the user-comments. Sentiment analysis, also known as opinion mining is the computational study of sentiments and opinions conveyed in natural language for the purpose of decision making. Preprocessing data play a vital role in getting accurate sentiment analysis results. Extracting opinion target words provide fine-grained analysis on the customer reviews. The labeled data required for training a classifier is expensive and hence to over come, Domain Adaptation technique is used. In this technique, Single classifier is designed to classify homogeneous and heterogeneous input from di_erent domain. Sentiment Dictionary used to find the opinion about a word need to be consistent and a number of techniques are used to check the consistency of the dictionaries. This paper focuses on the survey of the existing methods of Sentiment analysis and Opinion mining techniques from social media.

Keywords: domain adaptation; machine learning; opining mining; sentiment analysis; sentiment classification.

I. Introduction

Due to the huge growth of social media on the web, opinions extracted in these media are used by individuals and organizations for decision making. Each site contains a large amount of opinated text which makes it challenging for the user to read and extract information [1]. This problem can be overcome by using sentiment analysis techniques. The main objective of sentiment analysis is to mine sentiments and opinions expressed in the user generated reviews and classifing it into different polarities. The output is the data annotated with sentiment labels. Machine learning techniques are widely used for sentiment classification [2]. For a specific domain *D*, sentiment data *Xi* and *Yi* denoting data *Xi* has polarity *Yi*. If the overall sentiment expressed in *Xi* is positive, then *Yi* is +1, else -1. Labelled sentiment data is a pair of sentiment text and its corresponding sentiment polarity *fXi,Yig*. If *Xi* is not assigned with any polarity data *Yi*, then it is a unlabelled sentiment data. In supervised sentiment classi_cation method, classifier is trained using labeled data from a particular domain. Semisupervised classification method, combines unlabeled data along with few labeled training data to construct the classifier [3].

Applications: There are variety of information in the form of news blogs, twitter *etc..* are available in the social media about different products. Sentiment Analysis can summarize and give a score that represents the opinion of that data. This is used by customers depending on their need. There are a number of applications of sentiment analysis and opinion mining. The area where Sentiment Analysis is used is in Finance, Politics, Business and public actions. In business Domain, Sentiment analysis is used to detect the customer's interest in their product. Sentiment analysis in political do-main is used to get the clarity on the politician's position. Opinion Mining is also used to find the public interest on the newly applied rules by the goverenment.

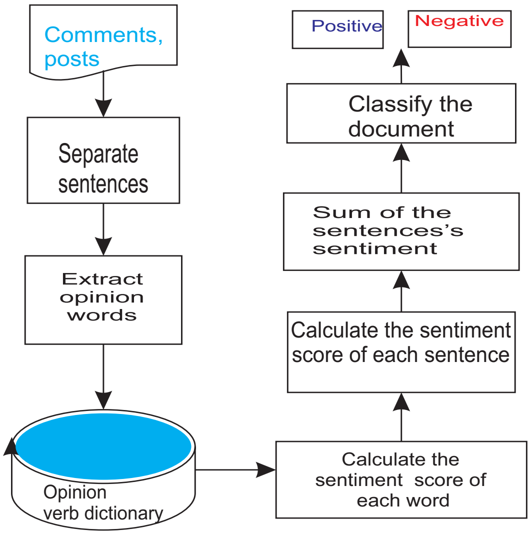

Comments, posts
Separate sentences
Extract opinion words
Opinion verb dictionary
Calculate the sentiment score of each word
Calculate the sentiment score of each sentence
Sum of the sentences's sentiment
Classify the document
Positive
Negative

Figure 1: Architecture of Sentiment Analysis

Motivation: Current trend is to look for opinions and sentiments in the product reviews that are available in large scale in social media. Before making decision, we tend to look at the sentiment analysis results of the opinion given by different users. This helps any customer to decide his opinion on that product. As data available in large scale, it is a laborious process to look

Author α σ ρ: Department of Computer Science and Engineering University Visvesvaraya College of Engineering, Bangalore University, Bengaluru, India. e-mail: hiremathsavitha@gmail.com

into all the user opinion. Hence Sentiment analysis is require. The main Objective of sentiment analysis is to classify the sentiment into different categories. Figure 1, shows the overall architecture of the sentiment analysis. Document level, sentence level and aspect level are the different levels of sentiment classification. Classifying each document into positive or negative class is called document-level sentiment classification. While expressing the sentiment of a document by this type of classifier, it assumes that document contains opinion of the user about a single object. Aspect level sentiment analysis classify the opinion about a document assuming that the opinion is expressed about different aspects in a document.

Sentiment classifiers, designed using data from one domain may not work with high accuracy if the same is used to classify the data from a different domain. One of the main reasons is that the sentiment words of a domain can be different from another domain. Thus, Domain adaptations are required to bridge the gaps between domains. The Domain used to train the classifier is called source domain and the domain to which we apply the trained classifier is called the target domain. The advantage of this method is that we need some or no labeled data of the target domain, where labeled data is costly as well as in-feasible to manually label the reviews for each domain type. This type of classification is called Cross Domain Sentiment Classification. Heterogeneous domain adaptation is required when domains of different dimension are input to the topic adaptive sentiment classifier.

Sentiment classifiers can be broadly classified into machine learning and lexicon based. Machine learning algorithms are used in machine learning approach. These algorithms can work in supervised, semi-supervised or unsupervised learning methods. Supervised learning methods give more accurate results compared to semi-supervised and unsupervised learning methods, but it requires labeled data which is expensive and time consuming process. Semi-supervised approach uses Easy Adapt (++[EA++]) which is easier than the Easy Adapt [EA] which requires labeled data from source and target domain. This is because it uses both labelled and unlabeled data from the target domain which results in superior performance theoretically and empirically over EA and hence it can be efficiently used for preprocessing [4]. Lexicon based approach utilizes Sentiment lexicon to analyze the sentiments in a review. Lexicon based approach can use dictionary or corpus to classify the sentiment words. Due to the shortage of labeled data, a single classifier can be designed to classify reviews from different domains. But classifier designed to classify data from one domain may not work efficiently on other domain. This is due to domain specific words which are different for every domain.

Support vector machine and Naive baye's classifiers are the important classifiers that support machine learning approach. Support vector machine classify data by finding hyper-plane that separates into different classes. Naive Baye's classifier is a probabilistic classifier based on Bayes theorem and the strong independence between the features. As there is a shortage of labeled data, a single classifier can be designed to classify reviews from different domains. But classifier designed to classify data from one domain may not work efficiently on other domain. This is due to domain specific words which are different for different domain.

Organization: The paper is organized as follows. Section 2 deals with the difffeerent techniques of data Preprocessing. In Section 3, Domain Adaptation Methods are discussed along with importance and applications of Heterogeneous Domain Adaptation. Section 4 give a comparison of different Topic Adaptive Sentiment Classi_cation methods. Sections 5 and 6 gives an overview of the Extracting Opinion Targets and Words and Different levels of Sentiment Analysis respectively. Section 7 gives a brief overview on how to work on inconsistent dictionaries. Dificulties and Solutions of the Polarity Shifting Detection are discussed in Section 8. Intrinsic and Extrinsic Domain Relevance is discussed in section 9. Section 10 contained information regarding Content-Based and Policy- Based Filtering policies. Section 11 brief about the Evaluation methods and paper is concluded in Section 12.

II. Preprocessing Data

Data provided in the form of reviews by the users contain lot of noise which need to be removed before it is classified. Haddi *et al.*[5] have explored the role of preprocessing in improving the SVM classifier results by selecting appropriate features. Selection of relevent features increase the accuracy of the classification process. Different techniques used are Feature Frequency, Term Frequency Inverse Document Frequency, Feature Presence. Boa *et al.* [6] show the effect of *urls*, repeated letters, negation, lemmatization and stemming on the performance of the classifier. Bigrams and emotion features addition improves the accuracy of the classifier [7].

There are mainly three steps in data processing, to-kenization, normalization and part-of-speech(POS) tagging. Transfering inected form to base form, also known as lemma is called lemmatization. This reduces the sparseness of the data which make text classification efficient [8]. Stemming processes a word without knowledge of the context. Whereas lemmatization considers contextual part-of-speech information while finding the base form of a word.

Unigrams and bigrams can be selected as training features. Pang *et al.* [9] show that unigrams turned out to be more effective compared to using bigrams. This leads to less features which give high performance. Stop words are excluded as they are not helpful for our classification.

III. Domain Adaptation Methods

Domain adaptation methods have been used for different research fields. According to the data in the target domain, the domain adaptation methods are generally divided into three categories: Supervised, Semi-supervised and Unsupervised domain adaptation methods. Supervised domain adaptation only use the labelled data in the target domain, Semisupervised domain adaptation use both the labelled and unlabelled data in the target domain and Unsupervised domain adaptation use only the unlabeled data in the target domain [4], [10]. Xavier *et al.* [11] proposed an efficient method for domain adaptation without the requirement of labeled data. This method classifies reviews from multiple domains by extracting the topic adaptive words from the unlabeled tweets using deep learning approach. SUIT model [12] considers the topic aspects and opinion holders for domain adaptation using supervised learning.

Daume *et al.* [18] proposed a feature augmentation method for domain adaptation. This method augments the source domain feature space using feature from labeled data from the target domain. Cheng *et al.* [17] proposed semi supervised domain adaptation method that maps source to target feature space. Methods proposed in [19] donot consider labeled data while considering learning feature representation. Ando *et al.* [20] proposed multitasking algorithm to select pivot features between source and target domains which is used to build pseudo-tasks for building correspondence among the features.

Structural Correspondence Learning uses unlabeled data from both source and target domain to obtain common features referred to as Pivots which behave in the same way in both domains and to find the correspondence between them. Non-pivot features which co-occur with pivot features are also considered. This technique is tested on the part of speech tagging and show the gain in performance for varying amount of source and target training data [21].

Table 1: Summary of the Survey of Domain Adaptation Techniques

Author	Concept	Advantages	Disadvantages
Bollegala *et al.*, (2016), [13]	Project both source and target domain in same lower dimensional embedding and then learning classifier on this embedded feature	Only source domain labeled data is used	Single rule is applied at a time
Liu *et al.*, (2015), [14]	Updates topic adaptive features based on collaborative selection of unlabelled data	Single classifier classifies multiple topic tweets	Few topic adaptive sentiment words are not selected due to the threshold applied while selecting the words
Quynh *et al.*, (2015), [15]	linguistic resources are used to generate additional training examples	Percentage of new training examples is high	Errors of syntactic parsing may cause problems
Xiao *et al.*, (2015), [16]	Feature Space Independent semi-supervised domain adaptation	Both Homogeneous and Heterogeneous domain adaptation is implemented	-
Cheng and Pan (2014), [17]	Linear Transformation from source to target damain is used with Semi-Supervised Adaptation	Method can be used in general for all variety of loss functions	Practical Domain Adaptation problems are not Considered

J Blitzer et al [22] proposed a method where the SCL algorithm is extended which reduces the error between the domains by 30 to 46 percent over supervised baseline. Movies reviews are the most studied domain in the early days, but at present the number of domains are increasing widely. The sentiment classification system has to collect data for each new domain. The pivots are selected not only by considering the frequency of occurrence but also by using the manual information of the source labels by using very small number of labeled information. The distance between the domains is obtained which is the measure of loss due to domain adaptation from one to another. Spectral Feature Alignment require only small amount of source domain labeled data and no label data from target domain. To span the gap between source and target domain spectral feature alignment algorithm is used to align the domain specific words into a unified cluster with the help of domain independent words as a bridge. SFA provides a new representation of cross-domain data by using the relationship between domain specific and domain independent features(pivots) by clustering them into the same latent space. These clusters reduce the mismatch between domain specific words of both domains.The classifier is trained on the new representation.

Bipartite graph is constructed to study the relationship between domain specific and domain independent words [23], [24]. A sentiment sensitive the saurus is created by using labeled data from diverse source domains and unlabeled data from both source and target domains to find the association between the words in different domains. The created thesaurus is used to expand the feature vector to train the binary classifier. The feature vector expansion is done by appending the additional features that represent the source and target domain reviews to minimize the mismatch of features [25], [13].

Locality Preserving Projections is a linear projective map that emerges by resolving the different problem and by maintaining the locality of the constitution of data set. When two data overlap on the other, with the decreasing dimensionalities in the ambient space the Locality Preserving Projections are derived by determining the optimal linear estimations to the eigen functions of the Laplace Beltrami operator. Training and trial data when drawn from same distribution methods of Discriminative learning execute well. Infinite number of labeled data are available for source domain, but, focus is to find a classifier that performs effectively on target with little or no labeled data. First, we have to evaluate the conditions on which the classifier performs well on the target domain. Second, having compact labeled data for target domain and huge labeled data for source domain we need to combine them during training to attain minimum mistakes at test time [26].

a) Heterogeneous Domain Adaptation

Domain adaptation methods assume that the data from different domains are represented by the same type of features with same dimensions. These methods cannot classify if the dimensions of source and target data are different. Technique of classifying such data is called Heterogeneous domain adaptation. Shi *et al.* [27] propose a solution where classification of high accuracy can be obtained even with the different feature space and different data distribution. Spectral embedding is used to unify the feature space of both source and target domains. It proposes Heterogeneous spectral mapping to find the common feature subspace by understanding two feature mapping matrix. Gap between the two domains in Domain adaptation methods can be achieved by re-weighting source instances [28], [29], target instances are self-labeled [30], [31], introducing new feature representations [22], [32], [33]. These methods can be applied when both domains have same feature representations. In real world, feature representation in the source domain can be completely different from target do-main while doing cross domain sentiment classification. Example for this is cross language text classification, where reviews from different language domains are represented by words in different languages. Text-aided image classification can also be executed where source domain has word features and target domain has visual features.

Number of approaches are employed for hetero geneous domain adaptation, such as heterogeneous spectral mapping [27], feature mapping, feature projection and transformation [34], [35], manifold alignment [36] and auxiliary resources [37]. Xiao *et al.* [16] proposed a method which can do homogeneous and heterogeneous domain adaptation across domains. In this process, source domain is assumed to have large set of labeled data and unlabeled data compared to target domain data. Instead of focusing on the feature divergence, each domain instances are employed kernelized representation. Table 1 gives the summary of the survey of various Domain Adaptation techniques.

IV. Topic Adaptive Sentiment Classification

Sentiment classifier trained using data from one domain may not give a good accuacy if the same classifier is used to classify data from different do-main. For example, sentiment words of kitchen do-main are different from book domain. Blitzer *et al.* [22] proposed an approach called structural correspondance learning for domain adaptation where it used pivot features to bridge the gap between source and target domain. Pan *et al.* [23] proposed a method called spectral feature alignment where domain specific words from different domains are aligned into unified clusters. Bollegala *et al.* [25] proposed a method for classification when we do not have labeled data of target domain, but we have few labeled data of other domains. This method automatically creates a sentiment sesitive Thesaurus using labeled and unlabeled data from multiple source domains. Constructed Thesaurus is then used to enlarge the feature vectors to train the classifier. Choi *et al.* [38] proposed linear integer programming method that can adapt an existing lexicon into a new one and find the relations among words and opinion expressions to find the most likely polarity of each lexicon item for the given domain.

Subjectivity analysis is concerned with extracting infor- mation about opinions, sentiments and other private states expressed in texts. Stoyanov *et al.* [39] proposed a method which collectively considers the relation among words and opinion statements to get the polarity of the sentiment words of the given domain. He *et al.* [40] and Gao *et al.* [41] gave a probabilistic topic model which bridge each pair of domains in a semantic level. Compared to review data, Twitter data contain more variety topics from various domains. To train a topic specific classifier, labeled data is required. Aspect level sentiment analysis detect topic, relation of topic aspects, opinion words and sentiment holders in a document [42], [43]. Supervised learning is used in SUIT model [44] considering topic aspects and opinion holders for cross domain sentiment classification. Mejova *et al.* [45], [46] have shown that by considering news, blogs and twitter data set, cross media sentiment classification can be done. Shenghua Liu *et al.* [14] proposed that a classifier designed using multiple domain twitter inputs can be used as a specific classifier to classify tweets from a specific domain. Microblogs as a social media has become an interesting input for sentiment analysis [47], [48], [49]. Tumasjan *et al.* [50] concluded that twitter messages are more oriented towards the political opinion. Supervised learning of a sentiment classifier need labeled tweets which is expensive and rarely available.

Semi-supervised Support vector machine is one of the efficient model which classify data with less labeled data and utilizing more unlabeled data. When features can be easily split into different views, co-training framework [51] achieves good results.

V. Extracting Opinion Targets and Opinion Words

Extracting opinion target and opinion words one of the important task of opinion mining. More attention has been given to focus on these tasks [52], [53]. Extraction can be classified into sentence level and corpus level extraction. Identifing opinion target/word in each sentence refers to sentence level extraction [54], [55]. Extractors such as CRFs and HMM are built using sequence labelling models. Huang [56] shows that opinion extraction can be done using lexicalized HMM model. These methods need labeled training data to train the model. Overall performance of extraction reduces if less amount of labeled training data or labelled data from different domains other than the current domain is used. Based on the transfer learning method Li *et al.* [57] proposed that Cross do-main sentiment extraction of opinion words/ targets. Performance of extraction depend more on the relevance between source and target domain.

Most of the earlier methods applied a unsupervised extraction process. Important component of this method is to detect opinion relations and finding opinion associations among the words. Hu *et al.* [58] show that nearest neighbour rule can also exploit opinion relations among words. To obtain the good accuracy of detecting opinion relations among the words, only considering nearest neighbour rule and co-occurrence information is not sufffficient. Specific patterns are designed by Zhang *et al.* and these are used in [59] which considerably increased recall. They also used HITS algorithm to calculate opinion target confidence to increase precision. Word Alignment Model is one of the important algorithm to extract opinion/target. Liu *et al.* [60] implemented WAM based opinion/target extraction. Thy used unsupervised WAM to capture opinion relations in sentences. From opinion relations, random walk framework is used to extract opinion targets. To detect implicit topics and opinon words, topic modeling is employed [61], [62]. The purpose of these method is not to extract opinion target/word, instead clustering all words with respect to the aspect in reviews.

VI. Sentiment Analysis at Different Levels

Sentiment analysis approaches extract sentiment words from the text and find the orientation of words to classify them as positive, negative or neutral words. Initially, sentiment analysis focused on the semantic orientation of adjectives. The techniques of analysing the sentiment words are largely used in filtering text, discovering the public opinions, customer relationship [63]. Sentiment analysis can be done at different levels of granularity from document level to sentence level. Pang *et al.* [64] proposed three machine learning algorithms: support vector machines, maximum entropy classification, and Naive Baye's give best results compared to human created baselines [64]. Rule based and Learning based approaches are the different categories of the sentiment analysis approaches. This approach uses the handbuild lexicon. Bloom *et al.* [65] propose a method that extracts the sentiment orientation from lexicon and classify the sentence or document by analysis the patterns that occure in text. Wiebe *et al.* [66] provided a lexicon containing subje-ctive words such as verbs, nouns, adjectives with their polarity and strength associated with them.

The polarity of word can change depending on the context in a sentence. Number of methods are proposed to find the sentiment orientation of words by considering the context of a sentence [67]. Yuen *et al.* gave an approach that calculate the sentiment orientation of words on the basis of morphemes and its statistical association with strong polarised words. To measure semantic polarity of adjectives, wordnet can also be used [68]. Hu and Liu [69] proposed a method where linguistic patterns called sequential rules are used to extract opinion features from reviews which can be mined from labeled data which is used for training sequences of words. Kim *et al.* [70] proposed a method for identifying an opinion with its holder and topic, given a sentence in on-line news media texts is extracted.

Millions of people daily post their comments on variety of topics with the help of social media. It is very difficult to analyse this information as it is huge and generally it is multidimentional and time varing. Wang *et al.* [71] proposed a visualization system that analysis the sentiments that are expressed in the public comments and give the short term trend of the sentiments. Relationship map is used to visualize the changes in the attributes and evolution model is used to compare the time varing parameters. In general, most of the machine learning algorithms learn single task at a time. Li*et al.* [72] investigated on Collaborative Multitasking Learning algorithm. The aim of the work is to focus on improving the performance of all tasks insted of single primary task. Online data is used for learning so that it can be processed as and when it arrives. This makes method more realistic. Collaborative Online Multitasking Learning algorithm results in improved classification performance. Relation between the tasks is assumed to be uniform and considering the relatedness degree among the tasks still improve the performance.

Social media is a great place for students to share their experiences, ideas, emotions, stress and to seek social support. To understand and reect the students experiences in the social media, human intervention is required. As the data available is huge, we need a automated classifier. To address this problem, Chen *et al.* [73] proposed a platform where Students Learning Experience is analysed by integrating large scale data mining techniques and Qualitative analysis. All students may not be active in social media, resulting in only few students who are ready to share their thoughts post their ideas. This work only focus on the text content, where as images and videos also add lot of information. Various research work is done on extracting sentiments from the comments the user has posted. Tan *et al.* [74] worked on finding the sentiment variations in the twitter that give insite about the reason behind the cause of sentiment variations. Latent Dirichlet Allocation model is used to analyse the possible reason for the sentiment variations.

VII. Polarity Inconsistent Dictionaries

Sentiment dictionaries are used to find the polarity of opinion words in the reviews. Orientation of opinion words in the reviews can be found using sentiment dictionaries. Final orientation of a sentence or a document is the addition of orientation of each word. There are different types of sentiment dictionaries. Domain independent sentiment dictionaries are created manually or semi automatically used by all domain reviews. Major problems with sentiment dictionaries is the inconsistency in intra and inter dictionary. Fragut *et al.* [75] show that inconsistency problem is NP complete. Inconsistency in dictionaries can be detected using fast SAT solver. There are corpora and Wordnet-based sentiment polarity lexicon used. To derive sentiment lexicon, Wordnet based approach uses lexicon relations defined in wordnet. Measuring the rela tive distance of a word from examples determine the sentiment of adjectives in Wordnet [76]. Synonyms and antonyms are used to increase the sets of words. One more method to increase the set of words by adding all synomyms of a polar word with polarity and antonyms with reverse polarity [77].When seed polar words are very few such as low resource language, method suffer from low recall [78]. In QW [79] synsets in word net are automatically anotated. If two synsets are assigned opposing polarites, then they are discarded. Machine learning algorithms as well as stochastic algorithms [80] can be used to classify words into different polarities.

VIII. Polarity Shifting Detection

Sentiment classifiers are intended to classify the document into different catagories. Bag of words model is used to represent the text which need to be classified. In BOW model, the text is represented in the form of vector of words. As BOW changes word order and remove some syntatic information, it is not an efficient model for sentiment classification. To remove this advantage, linguistic knowledge is introduced to enhance the efficiency of BOW. However, accuracy improvement is very less due to the basic aws in BOW. Polarity shift problem is the most important dificuly in the BOW model. Features are also used to determine whether the phrase is positive or negative contextual polarity and overall aim is to use the phrase-level sentiment analysis. Several approaches are proposed to address the polarity shift problem [81].

Polarity shift problem also has a problem of extra annotations and linguistic knowledge and some efforts are done on solving this problem [82], [83]. Nakagawa *et al.* [84] proposes a dependency tree-based method for Japanese and English sentiment classification using conditional random variables. The polarity of each dependency subtree of a subjective sentence is represented by a hidden variable. Sentiment classification is done by calculating the values of the hidden variables that are calculated in consideration of interaction between the variables. Liu *et al.* [85] proposes linguistic rules to deal with the problem together with a new option aggregation function and classifies the review or opinion whether it is a positive or negative. Ding *et al.* [86] came up with holistic lexicon based approach to determine the semantic orientation of the reviews obtained by opinion mining and uses a new function for aggregating multiple opinion words in the same sentence. Ding *et al.* [87] deals with the assigning of entities to the opinion extracted using a pattern based method. It also finds the entities of the comparative sentences whose entities are not explicitly mentioned by extracting large opinions using state of the art technique. Several techniques for opinion mining features based on data mining and Natural language processing (NLP) methods on product reviews. It gives a feature-based summary of a large number of product reviews by customer.

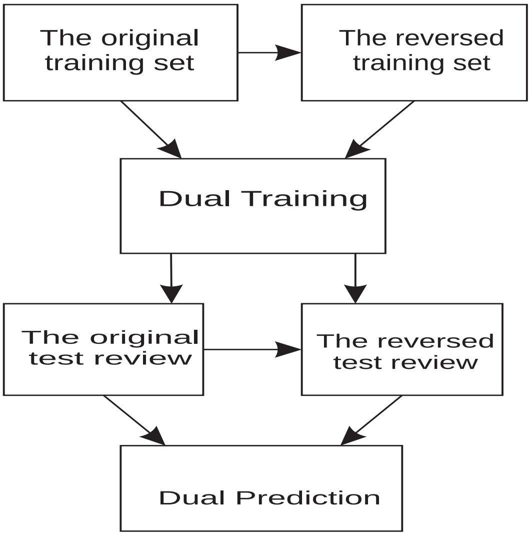


Figure 2: Dual sentiment analysis

Turney *et al.* [88] proposed a concept based on simple unsupervised learning algorithm for rating a review as recommended or un-recommended. The algorithm extract phrases which has adjectives or adverbs and estimates the semantic orientation of each phrase and classifies the reviews based on average semantic orientation. Turney *et al.* [89] provides general technique to measure semantic orientation to semantic association. It evaluates the semantic orientation using Pointwise mutual information and Latentsentiment Analysis(LSA) methods. Determining the polarity of sentiment-bearing expression, by considering the effect of interaction among words or constituents is important. It provides novel training-based approach which incorporates the structural inference to the learning procedure by the compositional semantics [90].

a) *Data Expansion Technique*

Expanding the data has been seen in the handwritten recognition, where the performance of this method is improved by adding few more training data. Figure 2 gives the process for dual sentiment analysis. In text mining, Agirre *et al.* [91] proposed a method to expand the amount of labeled data unique expressions in definitions from wordnet for a task of word sense disambiguation. Fujita *et al.* [92] proposed a method which provides training data using sentences from the external dictionary. Xia *et al.* [93] proposed a novel method of data expansion. The original and reversed reviews are constructed in one to one correspondence. The data expansion happens both in training stage and also during testing stage.

IX. Intrinsic and Extrinsic Domain Relevance

Opinion feature indicate attribute of an entity or an entity on which user express their opinions. Many approaches are proposed to extract to classify movie review opinion features. One of the efficient method is supervised learning method. This method works well in a given domain and if it needs to work for other domain, it has to be retrained [94].

By defining domain independent syntactic rules, Unsupervised approaches [95] identify opinion features. Wiebe *et al.* [96] proposed a supervised classification method to predict sentence subjectivity [97]. Pang *et al.* [98] proposed three machine learning algorithms to classify movie reviews into different sentiments. They are Naive Baye's, Support vector machines and maximum entropy [99].

A document can contain both Subjective and objective sentences. Due to this, Sentiment classifier may consider irrelevent text. Pang *et al.*, [100] proposed sentiment level subjectivity detector which identifies subjective or objective sentences. Then objective sentences are discarded which improves the classification results. Subjective sentences are further classified into positive and negative [101].

Wiebe *et al.* [102] proposed a method which uses naive Bayesian classifier to classify subjective sentences. One of the restriction for this method is the shortage of training set. Riloff *et al.* [103] proposed bootstrapping method which automatically label the training data so that lack of training data problem is solved.

X. Information Filtering

Online social networks has become a popular interactive medium to communicate between the users.Every day there is exchange of huge amountof information between the users. Information may be text, audio and video data. But the disadvantage is user wall is posted with so many different varieties of information in which the user may not be interested in some perticular type of data. This leads to the requirement of filtering the messages on the user wall before posting it [104], [105]. User is given the authority to decide which content type of messages need to be blocked. Information filtering of textual documents is of a great concern in recent years [106]. Vanetti *et al.* [107] proposed a Filtered Wall(FW), an automated system which is able to filter unwanted messages from online social network users. To mechanically assign with every text messages a set of categaries based on content, Machine learning text categorization techniques [107] are used.

a) *Content-Based Filtering*

Information filtering system are used to classify continuously generated messages sent by information produces and post messages on to the user wall that may satisfy the user requirements. Content based filtering system selects messages based on the interrelationship between the contents of the messages and the user preferences. Content-based filtering system mainly use the machine learning algorithms. Here classifier is trained by learning from the labeled examples. Text is mapped into a condensed representation of its content and then applied to training by feature extraction procedure. Hirsh *et al.*, [108] improved the short text messages using semisupervised learning strategy. It is based on the combination of labeled training data and secondary corpus of unlabeled data. Another approach proposed by Bobicev *et al.*, [109] is to adapt a statistical learning method that performs well.

b) *Policy based Personalization*

Classification mechanisms for personalizing access in OSNs is of recent interest. In [110], focus is on twitter and each tweet is associated with set of categories depending on its content. User selects tweets depending on the content that they are interested in. Contradicting to this, Golbeck *et al.*, [111] proposed *filmTrust*, that gives OSN trust relationship and this does not provide filtering policy layer by layer. Hence, user cannot exploit the classification results.

c) *Text Representation*

To increase the performance of classifier, taking out suitable set of features which present the text of a document is necessary. There are divergent sets of features for text classification. BOW, Document properties (Dp) and Contextual features(CF) [112], [113] are considered for short text messages. BOW and DP are used in [112] and they are completely derived from the information present with in the text of the message. Contextual features play an important role in finding the semantics of the messages.

XI. Evaluation

The performance of variety of methods that are used in sentiment analysis is compared by measuring few parameters like precision, recall and Fscore. Precision is a part of retrieved data that are more applicable. Whereas recall is the part of relevant data that are retrieved. F-measure is calculated using both recall and precision. As given in the Table 2, we have compared various works with respect to classifiers used, feature extraction methods and different measurement metrics. Matrics considered are Accuracy(A), Precision(P), Recall(R) and F-score(F).

XII. Conclusions

Variety of applications of sentiment analysis are widely used. They include classifing reviews, sum marizing review etc. In this paper, we have discussed different approaches of sentiment classification and its performance. Domain adaptation is required as it reduces number of classifier user for the sentiment analysis. Different approaches of domain adaptation are compared using supervised, semisupervised and unsupervised learning methods. Heterogeneous domain adaptation is able to classify data with different dimensions. Extracting opinion words and target words is crucial for the performance of the classifier. Efficient algorithms for extracting opinions words and opinion target are discussed. Data expansion techniques are discussed which is used in dual sentiment analysis that reduces the number of training labeled data used for the classification. Information filtering is a online social network user friendly concept which gives user exible choices.

References Références Referencias

1. G. U. Vasanthakumar, P. D. Shenoy, and K. R. Venugopal, PTIB: Profiling Top Inuential Blogger in Online Social Networks," vol. 10, no. 1, pp. 77-91, 2016.
2. N. Godbole, M. Srinivasaiah, and S. Skiena, Large-Scale Sentiment Analysis for News and Blogs." *ICWSM*, vol. 7, no. 21, pp. 219 -222, 2007.
3. S. Li, C.-R. Huang, G. Zhou, and S. Y. M. Lee, Employing Personal / impersonal Views in Supervised and Semi-supervised Sentiment Classification," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 414-423, 2010.
4. J. Jiang and C. Zhai, Instance Weighting for Domain Adaptation in NLP," in *Proceedings of the 45th*

Annual meeting of the Association for Computational Linguistics, vol. 7, pp. 264-271, 2007.

5. E. Haddi, X. Liu, and Y. Shi, The Role of Text Pre-processing in Sentiment Analysis," *Proceedings of the First International Conference on Information Technology and Quantitative Management, Procedia Computer Science*, vol. 17, pp. 26 -32, 2013.
6. Y. Bao, C. Quan, L. Wang, and F. Ren, The Role of Pre-processing in Twitter Sentiment Analysis," in *Proceedings of the International Conference on Intelligent Computing Methodologies*. Springer, pp. 615 - 624, 2014.
7. K. R. Venugopal, K. G. Srinivasa, and L. M. Patnaik, *Soft Computing for Data Mining Applications*. Springer, 2009.
8. D. Sejal, K. G. Shailesh, V. Tejaswi, D. Anvekar, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, “Query Click and Text Similarity Graph for Query Suggestions," in *Proceedings of the International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, pp. 328{341, 2015.
9. B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment Classification using Machine Learning Techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Languag Processing, ACL*, vol. 10, pp. 79-86, 2002.
10. D. Davidov, O. Tsur, and A. Report, Semi-supervised Recognition of Sarcastic Sentences in Twitter and Amazon," in *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 107-116, 2010.
11. X. Glorot, A. Bordes, and Y. Bengio, Domain Adaptation for Large-scale Sentiment Classification: A Deep Learning Approach," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 513-520, 2011.
12. F. Li, S. Wang, S. Liu, and M. Zhang, Suit: A Supervised User-item Based Topic Model for Sentiment Analysis," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1636-1642, 2014.
13. D. Bollegala, T. Mu, and J. Goulermas, Cross-domain Sentiment Classification using Sentiment Sensitive Embeddings," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 2, pp. 398 - 410, 2016.
14. S. Liu, X. Cheng, F. Li, and F. Li, TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 6, pp. 1696 -1709, 2015.

Table 2: Comparison of the Various Sentiment Analysis Techniques along with Performance

Author	Concept	Feature selection method	Data source	A	P	R	F
Shenghua Liu *et al.*, (2015), [14]	Multiclass SVM Classifier	PMI-IR	Sander-Twitter Corpus	0.54	0.55	0.52	0.53
Xiao *et al.*, (2015), [16]	Hilbert Schmidt Independence Criteria	Prediction function	Amazon product reviews	0.79	-	-	-
Rui Xia *et al.*, (2015), [93]	Naive bayes classifier	Data Expansion Technique	Amazon product reviews	0.81	-	-	-
Wen li *et al.*, (2014), [114]	SVM Classifier	Heterogeneous feature augmentation	Object dataset	0.54	-	-	-
Jianping cao *et al.*, (2014), [115]	Rule based classifier	Sentiment polarity score	tianya.cn	0.82	0.84	0.30	-
Li Cheng *et al.*, (2014), [17]	SVM classifer	Semi-supervised transfer component analysis	Amazon product reviews	0.63	-	-	-
Zhen Hai *et al.*, (2014), [116]	Differentiating opinion feature statistics across domains	Intrinsic and Extrinsic domain relevance	Cell phone review corpus	-	0.65	0.61	0.63

15. M.-F. M. Quynh Thi Ngoc Do, Steven Bethard, Domain Adaptation in Semantic Role Labeling Using a Neural Language Model and Linguistic Resources," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1812-1823, 2015.
16. M. Xiao and Y. Guo, Feature Space Independent Semi-supervised Domain Adaptation *via* Kernel Matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 54 -66, 2015.
17. L. Cheng and S. J. Pan, Semi-Supervised Domain Adaptation on Manifolds," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 12, pp. 2240 - 2249, 2014.

18. H. Daum'e III, Frustratingly Easy Domain Adaptation," *arXiv preprint arXiv: 0907.1815*, 2009.
19. S. J. Pan, J. T. Kwok, and Q. Yang, Transfer Learning *via* Dimensionality Reduction." in *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, vol. 8, pp. 677 -682, 2008.
20. R. K. Ando and T. Zhang, A Framework for Learning Predictive Structures from Multiple Tasks and Unlabeled Data," *The Journal of Machine Learning Research*, vol. 6, no. 2, pp. 1817-1853, 2005.
21. J. Blitzer, R. McDonald, and F. Pereira, Domain Adaptation with Structural Correspondence Learning," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp.120-128, 2006.
22. J. Blitzer, M. Dredze, F. Pereira *et al.*, Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification," in *Proceedings of the 45th Annual Meeting of the Association of the Computational Linguistics, ACL*, vol. 7, pp. 440 - 447, 2007.
23. S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen, Cross-domain Sentiment Classification *via* Spectral Feature Alignment," in *Proceedings of the 19th International Conference on World wide web*. ACM, pp. 751-760, 2010.
24. G. Vasanthakumar, P. D. Shenoy, and K. R. Venugopal, PFU: Profiling Forum Users in Oline Social Networks, a Knowledge Driven Data Mining Approach," in *Proceedings of the IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, pp. 57- 60, 2015.
25. D. Bollegala, D. Weir, and J. Carroll, Using Multiple Sources to Construct a Sentiment Sensitive The saurus for Cross-domain Sentiment Classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 132-141, 2011.
26. S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, A Theory of Learning from Different Domains," *Machine Learning*, vol. 79, no. 1, pp. 151-175, 2010.
27. X. Shi, Q. Liu, W. Fan, P. S. Yu, and R. Zhu, Transfer Learning on Feature Spaces *via* Spectral Transformation," in *Proceedings of the IEEE 10th International Conference on Data Mining (ICDM)*. IEEE, pp. 1049 -1054, 2010.
28. M. Sugiyama, S. Nakajima, H. Kashima, P. V. Buenau, and M. Kawanabe, Direct Importance Estimation with Model Selection and its Application to Covariate Shift Adaptation," in *Proceedings of the Advances in Neural Information Processing Systems*, pp1433-1440, 2008.
29. H. Shimodaira, Improving Predictive Inference unde Covariate Shift by Weighting the Log-likelihood Function," *Journal of Statistical Planning and Inferen-*cevol. 90, no. 2, pp. 227-244, 2000.
30. M. Chen, K. Q. Weinberger, and J. Blitzer, Cotraining for Domain Adaptation," in *Proceedings othe Advances in Neural Information Processing Sytems*, pp. 2456-2464, 2011
31. G. Tur, Co-adaptation: Adaptive Co-training for Semi-supervised Learning," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009.* IEEE, pp. 3721-3724, 2009.
32. Kumar, A. Saha, and H. Daume, Coregularization based Semi-supervised Domain Adaptation," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 478-486, 2010.
33. J. Blitzer, S. Kakade, and D. P. Foster, Domai Adaptation with Coupled Subspaces," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 173 -181, 2011.
34. K. Saenko, B. Kulis, M. Fritz, and T. Darrell, Adapting Visual Category Models to New Domains," in *Proceedings of the European Conference on Compute Vision -ECCV*. Springer, pp. 213{226, 2010.
35. B. Kulis, K. Saenko, and T. Darrell, What You Saw is not What You Get: Domain Adaptation using Asymmetric Kernel Transforms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 1785-1792, 2011.
36. C. Wang and S. Mahadevan, Heterogeneous Domain Adaptation using Manifold Alignment," in *Procee dings of the International Joint Conference on Artifi cial Intelligence, IJCAI*, vol. 22, no. 1, pp. 1541, 2011.
37. W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, Translated Learning: Transfer Learning across Different Feature Spaces," in *Proceedings of the Advances in Neural Information Processing Systems*, pp. 353 - 360, 2008.
38. Y. Choi and C. Cardie, Adapting a Polarity Lexicon using Integer Linear Programming for Domain Specific Sentiment Classification," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, ACL*, vol. 2, pp. 590 - 598, 2009.
39. V. Stoyanov and C. Cardie, Topic Identification for Fine-grained Opinion Analysis," in *Proceedings of the 22nd International Conference on Computational Linguistics, Association for Computational Linguistics*, vol. 1, pp. 817- 824, 2008.
40. Y. He, C. Lin, and H. Alani, Automatically Extracting Polarity-bearing Topics for Cross-domain Sentiment Classification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pp. 123-131, 2011.
41. S. Gao and H. Li, A Cross-domain Aadaptation Method for Sentiment Classification using Probabilistic Latent Analysis," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. ACM, pp. 1047{1052, 2011.

42. V. Stoyanov and C. Cardie, Topic Identification for Fine-grained Opinion Analysis," in *Proceedings of the 22nd International Conference on Computational Linguistics, Association for Computational Linguistics*, vol. 1, pp. 817-824, 2008.
43. D. Das and S. Bandyopadhyay, Emotion Coreferencing-Emotional Expression, Holder, and Topic," *Computational Linguistics and Chinese Language Processing*, vol. 18, no. 1, pp. 79-98, 2013.
44. F. Li, S. Wang, S. Liu, and M. Zhang, Suit: A Supervised User-item based Topic Model for Sentiment Analysis," in *Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 1636-1642, 2014.
45. Y. Mejova and P. Srinivasan, Crossing Media Streams with Sentiment: Domain Adaptation in Blogs, Reviews and Twitter." in *Sixth International AAAI Conference on Weblog and Social Media*, pp. 234 - 241, 2012.
46. D. Das and S. Bandyopadhyay, Extracting Emotion Topics from Blog Sentences: Use of Voting from Multiengine Supervised Classifiers," in *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. ACM, pp. 119-126, 2010.
47. S. Liu, F. Li, F. Li, X. Cheng, and H. Shen, Adaptive Cotraining SVM for Sentiment Classification on Tweets," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. ACM, pp. 2079 - 2088, 2013.
48. A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, Sentiment Analysis of Twitter Data," in *Proceedings of the Workshop on Languages in Social Media*. Association for Computational Linguistics, pp. 30 - 38, 2011.
49. S. Liu, W. Zhu, N. Xu, F. Li, X.-q. Cheng, Y. Liu, and Y. Wang, Co-training and Visualizing Sentiment Evolvement for Tweet Events," in *Proceedings of the 22nd International Conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, pp. 105-106, 2013.
50. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, vol. 10, pp. 178 -185, 2010.
51. D. Das and S. Bandyopadhyay, Identifying Emotion Topic An Unsupervised Hybrid Approach with Rhetorical Structure and Heuristic Classifier," in *Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*. IEEE, pp. 1-8, 2010.
52. Y. Kim and S. R. Jeong, Opinion-Mining Methodology for Social Media Analytics." *TIIS*, vol. 9, no. 1, pp. 391-406, 2015.
53. H. Yu, Structure-aware Review Mining and Summarization," in *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 653 - 661, 2010.
54. T. Ma and X. Wan, Opinion Target Extraction in Chinese News Comments," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pp. 782-790, 2010.
55. Q. Zhang, Y. Wu, T. Li, M. Ogihara, J. Johnson, and X. Huang, Mining Product Reviews Based on Shallow Dependency Parsing," in *Proceedings of the 32rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 726-727, 2009.
56. W. Jin, H. H. Ho, and R. K. Srihari, A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining," in *Proceedings of the 26th Annual International Conference on Machine Learning*. Citeseer, pp. 465 -472, 2009.
57. F. Li, S. J. Pan, O. Jin, Q. Yang, and X. Zhu, Cross-domain Co-extraction of Sentiment and Topic Lexicons," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, vol. 1, pp. 410-419, 2012.
58. M. Hu and B. Liu, Mining Opinion Features in Customer Reviews," in *AAAI*, vol. 4, no. 4, pp. 755-760, 2004.
59. G. Qiu, B. Liu, J. Bu, and C. Chen, Opinion Word Expansion and Target Extraction Through Double Propagation," *Computational Linguistics*, vol. 37, no. 1, pp. 9-27, 2011.
60. K. Liu, L. Xu, and J. Zhao, Opinion Target Extraction using Word-based Translation Model," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1346 -1356, 2012.
61. W. X. Zhao, J. Jiang, H. Yan, and X. Li, Jointly Modeling Aspects and Opinions with a Maxent-LDA hybrid," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 56-65, 2010.
62. Mukherjee and B. Liu, Modeling Review Comments," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers*, vol. 1, pp. 320-329, 2012.
63. V. Jha, N. Manjunath, P. D. Shenoy, K. Venugopal, and L. Patnaik, HOMS: Hindi Opinion Mining System," in *IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS)*. IEEE, pp. 366 - 371, 2015.
64. B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment Classification using Machine Learning Techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, ACL*, vol. 10, pp. 79-86, 2002.
65. K. Bloom, N. Garg, and S. Argamon, Extracting Appraisal Expressions," in *HLT-NAACL*, pp. 308-315, 2007.

66. J. Wiebe, T. Wilson, R. Bruce, M. Bell, and M. Martin, Learning Subjective Language," *Computational linguistics*, vol. 30, no. 3, pp. 277-308, 2004.

67. L. Vibha, G. Harshavardhan, K. Pranaw, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, Classification of Mammograms using Decision Trees," in *Proceedings of the 10th International Database Engineering and Applications Symposium (IDEAS'06)*. IEEE, pp. 263 -266, 2006.

68. G. Vasanthakumar, A. K. Upadhyay, P. F. Kalmath, S. Dinakar, P. D. Shenoy, and K. Venugopal, UP3: User Profiling from Profile Picture in Multi-Social Networking," in *Proceedings of the Annual IEEE India Conference (INDICON), IEEE*, pp. 1-6, 2015.

69. M. Hu and B. Liu, Opinion Feature Extraction Using Class Sequential Rules." in *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pp. 61- 66, 2006.

70. S.-M. Kim and E. Hovy, Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text," in *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. ACL, pp. 1-8, 2006.

71. C. Wang, Z. Xiao, Y. Liu, Y. Xu, A. Zhou, and K. Zhang, SentiView: Sentiment Analysis and Visualization for Internet Popular Topics," *IEEE transactions on human-machine systems*, vol. 43, no. 6, pp. 620-630, 2013.

72. G. Li, S. C. Hoi, K. Chang, W. Liu, and R. Jain, Collaborative Online Multitask Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1866-1876, 2014.

73. X. Chen, M. Vorvoreanu, and K. Madhavan, Mining Social Media Data for Understanding Students' Learning Experiences," *IEEE Transactions on Learning Technologies*, vol. 7, no. 3, pp. 246-259, 2014.

74. S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, Interpreting the Public Sentiment Variations on Twitter," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 5, pp. 1158 - 1170, 2014.

75. E. Dragut, H. Wang, C. Yu, P. Sistla, and W. Meng, Polarity Consistency Checking for Sentiment Dictionaries," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 997-1005, 2012.

76. G. Vasanthakumar, B. Prajakta, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, PIB: Profiling Inuential Blogger in Online Social Networks, A Knowledge Driven Data Mining Approach," *Procedia Computer Science*, vol. 54, pp. 362-370, 2015.

77. S.-M. Kim and E. Hovy, Identifying and analyzing judgment opinions," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. ACL, pp. 200-207, 2006.

78. D. Rao and D. Ravichandran, Semi-supervised Polarity Lexicon Induction," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 675-682, 2009.

79. V. Jha, R. Savitha, S. Hebbar, P. D. Shenoy, and K. R. Venugopal, HMDSAD: Hindi Multi-domain Sentiment Aware Dictionary," in *Proceedings of the International Conference on Computing and Network Communications (CoCoNet), IEEE*, pp. 241-247, 2015.

80. A. Hassan and D. Radev, Identifying Text Polarity using Random Walks," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. ACL, pp. 395-403, 2010.

81. Kennedy and D. Inkpen, Sentiment Classification of Movie Reviews using Contextual Valence Shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110-125, 2006.

82. S. Li, S. Y. M. Lee, Y. Chen, C.-R. Huang, and G. Zhou, Sentiment Classification and Polarity Shifting," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, pp. 635 - 643, 2010.

83. T. Wilson, J. Wiebe, and P. Hoffmann, Recognizing Contextual Polarity: An Exploration of Features for Phrase-level Sentiment Analysis," *Computational Linguistics*, vol. 35, no. 3, pp. 399-433, 2009.

84. T. Nakagawa, K. Inui, and S. Kurohashi, Dependency Tree-based Sentiment Classification using CRF s with Hidden Variables," in *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 786 - 794, 2010.

85. X. Ding and B. Liu, The Utility of Linguistic Rules in Opinion Mining," in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 811- 812, 2007.

86. X. Ding, B. Liu, and P. S. Yu, A Holistic Lexicon-based Approach to Opinion Mining," in *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, pp. 231-240, 2008.

87. X. Ding, B. Liu, and L. Zhang, Entity Discovery and Assignment for Opinion Mining Applications," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1125-1134, 2009.

88. P. D. Turney, Thumbs Up or Thumbs Down? : Semantic Orientation Applied to Unsupervised Classification of Reviews," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 417-424, 2002.

89. P. D. Turney and M. L. Littman, Measuring Praise and Criticism: Inference of Semantic Orientation

from Association," *ACM Transactions on Information Systems (TOIS)*, vol. 21, no. 4, pp. 315-346, 2003.

90. Y. Choi and C. Cardie, Learning with Compositional Semantics as Structural Inference for Sub sentential Sentiment Analysis," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 793-801, 2008.
91. E. Agirre and D. Martinez, Exploring Automatic Word Sense Disambiguation with Decision Lists and the Web," in *Proceedings of the COLING-2000 Work shop on Semantic Annotation and Intelligent Content*. Association for Computational Linguistics, pp. 11 -19, 2000.
92. S. Fujita and A. Fujino, Word Sense Disambiguation by Combining Labeled Data Expansion and Semi-supervised Learning Method," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 12, no. 2, p. 7, 2013.
93. R. Xia, F. Xu, C. Zong, Q. Li, Y. Qi, and T. Li, Dual Sentiment Analysis: Considering Two Sides of One Review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 8, pp. 2120-2133, 2015.
94. N. Jakob and I. Gurevych, Extracting Opinion Targets in a Single-and Cross-domain Setting with Conditional Random Fields," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1035-1045, 2010.
95. G. Qiu, B. Liu, J. Bu, and C. Chen, Opinion Word Expansion and Target Extraction through Double Propagation," *Computational linguistics*, vol. 37, no. 1, pp. 9-27, 2011.
96. V. Hatzivassiloglou and J. M. Wiebe, Effects of Adjective Orientation and Gradability on Sentence Subjectivity," in *Proceedings of the 18th Conference on Computational linguistics, ACL*, vol. 1, pp. 299-305, 2000.
97. V. Jha, N. Manjunath, P. D. Shenoy, and K. R. Venugopal, HSAS: Hindi Subjectivity Analysis System," in *Proceedings of the 2015 Annual IEEE India Conference (INDICON)*. IEEE, pp. 1-6, 2015.
98. B. Pang, L. Lee, and S. Vaithyanathan, Thumbs up?: Sentiment Classification using Machine Learn- ing Techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, ACL*, vol. 10, pp. 79-86, 2002.
99. L. Vibha, G. Harshavardhan, K. Pranaw, P. D. Shenoy, K. R. Venugopal, and L. M. Patnaik, Lesion Detection using Segmentation and Classification of Mammograms," in *Proceedings of the 25th ASTED International Multi-Conference: Artificial Intelligence and Applications*. ACTA Press, pp. 311-316, 2007.
100. B. Pang and L. Lee, A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts," in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pp. 271, 2004.
101. B. Liu, Sentiment Analysis and Subjectivity." *and-book of Natural Language Processing*, vol. 2, pp. 627- 666, 2010.
102. J. M. Wiebe, R. F. Bruce, and T. P. O'Hara, Development and Use of a Gold-standard Data Set for Subjectivity Classifications," in *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, pp. 246- 253, 1999.
103. E. Riloff and J. Wiebe, Learning Extraction Pat- terns for Subjective Expressions," in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 105 -112, 2003.
104. K. Srinivasa, A. Singh, A. Thomas, K. R. Venugopal, and L. Patnaik, Generic Feature Extraction for Classification using Fuzzy C-means Clustering," in *Proceedings of the 3rd International Conference on Intelligent Sensing and Information Processing*. IEEE, pp. 33-38, 2005.
105. D. Sejal, V. Rashmi, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, Image Recommendation based on Keyword Relevance using Absorbing Markov Chain and Image Features," *International Journal of Multimedia Information Retrieval*, vol. 5, no. 3, pp. 185 - 199, 2016.
106. G. Adomavicius and A. Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engi- neering*, vol. 17, no. 6, pp. 734 -749, 2005.
107. M. Vanetti, E. Binaghi, E. Ferrari, B. Carminati, and M. Carullo, A System to Filter Unwanted Messages from OSN User Walls," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 2, pp. 285 - 297, 2013.
108. S. Zelikovitz and H. Hirsh, Improving Short Text Classification using Unlabeled Background Knowledge to Assess Document Similarity," in *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 1183-1190, 2000.
109. V. Bobicev and M. Sokolova, An Effective and Robust Method for Short Text Classification." in *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 1444 -1445, 2008.
110. B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, Short Text Classification in Twitter to Improve Information Filtering," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 841-842, 2010.
111. J. Golbeck, Combining Provenance with Trust in Social Networks for Semantic web Content Filtering," in *Provenance and Annotation of Data*. Springer, pp. 101-108, 2006.

112. M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, Content-based Filtering in On-line Social Networks," in *Privacy and Security Issues in Data Mining and Machine Learning*. Springer, pp. 127-140, 2010.
113. M. Carullo, E. Binaghi, and I. Gallo, An Online Document Clustering Technique for Short Web Contents," *Pattern Recognition Letters*, vol. 30, no. 10, pp. 870 - 876, 2009.
114. L. Duan, D. Xu, and I. Tsang, Learning with Augmented Features for Heterogeneous Domain Adaptation," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1134-1148, 2014.
115. J. Cao, K. Zeng, H. Wang, J. Cheng, F. Qiao, D. Wen, and Y. Gao, Web-Based Traffic Sentiment Analysis: Methods and Applications," *IEEE transactions on Intelligent Transportation systems*, vol. 15, no. 2, pp. 844-853, 2014.
116. Z. Hai, K. Chang, J.-J. Kim, and C. C. Yang, Identifying Features in Opinion Mining *via* Intrinsic and Extrinsic Domain Relevance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 3, pp. 623-634, 2014.

Global Journals Inc. (US) Guidelines Handbook 2016

www.GlobalJournals.org

Fellows

FELLOW OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (FARSC)

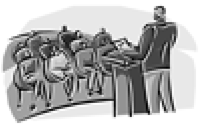
Global Journals Incorporate (USA) is accredited by Open Association of Research Society (OARS), U.S.A and in turn, awards “FARSC” title to individuals. The 'FARSC' title is accorded to a selected professional after the approval of the Editor-in-Chief/Editorial Board Members/Dean.

- The “FARSC” is a dignified title which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

FARSC accrediting is an honor. It authenticates your research activities. After recognition as FARSC, you can add 'FARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, and Visiting Card etc.

The following benefits can be availed by you only for next three years from the date of certification:

FARSC designated members are entitled to avail a 40% discount while publishing their research papers (of a single author) with Global Journals Incorporation (USA), if the same is accepted by Editorial Board/Peer Reviewers. If you are a main author or co-author in case of multiple authors, you will be entitled to avail discount of 10%.



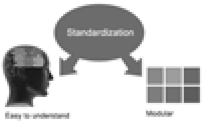
Once FARSC title is accorded, the Fellow is authorized to organize a symposium/seminar/conference on behalf of Global Journal Incorporation (USA).The Fellow can also participate in conference/seminar/symposium organized by another institution as representative of Global Journal. In both the cases, it is mandatory for him to discuss with us and obtain our consent.

You may join as member of the Editorial Board of Global Journals Incorporation (USA) after successful completion of three years as Fellow and as Peer Reviewer. In addition, it is also desirable that you should organize seminar/symposium/conference at least once.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The FARSC can go through standards of OARS. You can also play vital role if you have any suggestions so that proper amendment can take place to improve the same for the benefit of entire research community.

As FARSC, you will be given a renowned, secure and free professional email address with 100 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.



The FARSC will be eligible for a free application of standardization of their researches. Standardization of research will be subject to acceptability within stipulated norms as the next step after publishing in a journal. We shall depute a team of specialized research professionals who will render their services for elevating your researches to next higher level, which is worldwide open standardization.

The FARSC member can apply for grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A. Once you are designated as FARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria. After certification of all your credentials by OARS, they will be published on your Fellow Profile link on website https://associationofresearch.org which will be helpful to upgrade the dignity.

The FARSC members can avail the benefits of free research podcasting in Global Research Radio with their research documents. After publishing the work, (including published elsewhere worldwide with proper authorization) you can upload your research paper with your recorded voice or you can utilize chargeable services of our professional RJs to record your paper in their voice on request.

The FARSC member also entitled to get the benefits of free research podcasting of their research documents through video clips. We can also streamline your conference videos and display your slides/ online slides and online research video clips at reasonable charges, on request.

The FARSC is eligible to earn from sales proceeds of his/her researches/reference/review Books or literature, while publishing with Global Journals. The FARSC can decide whether he/she would like to publish his/her research in a closed manner. In this case, whenever readers purchase that individual research paper for reading, maximum 60% of its profit earned as royalty by Global Journals, will be credited to his/her bank account. The entire entitled amount will be credited to his/her bank account exceeding limit of minimum fixed balance. There is no minimum time limit for collection. The FARSC member can decide its price and we can help in making the right decision.

The FARSC member is eligible to join as a paid peer reviewer at Global Journals Incorporation (USA) and can get remuneration of 15% of author fees, taken from the author of a respective paper. After reviewing 5 or more papers you can request to transfer the amount to your bank account.

MEMBER OF ASSOCIATION OF RESEARCH SOCIETY IN COMPUTING (MARSC)

The ' MARSC ' title is accorded to a selected professional after the approval of the Editor-in-Chief / Editorial Board Members/Dean.
The “MARSC” is a dignified ornament which is accorded to a person’s name viz. Dr. John E. Hall, Ph.D., MARSC or William Walldroff, M.S., MARSC.

MARSC accrediting is an honor. It authenticates your research activities. After becoming MARSC, you can add 'MARSC' title with your name as you use this recognition as additional suffix to your status. This will definitely enhance and add more value and repute to your name. You may use it on your professional Counseling Materials such as CV, Resume, Visiting Card and Name Plate etc.

The following benefitscan be availed by you only for next three years from the date of certification.

MARSC designated members are entitled to avail a 25% discount while publishing their research papers (of a single author) in Global Journals Inc., if the same is accepted by our Editorial Board and Peer Reviewers. If you are a main author or co-author of a group of authors, you will get discount of 10%.

As MARSC, you will be given a renowned, secure and free professional email address with 30 GB of space e.g. johnhall@globaljournals.org. This will include Webmail, Spam Assassin, Email Forwarders, Auto-Responders, Email Delivery Route tracing, etc.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.

The MARSC member can apply for approval, grading and certification of standards of their educational and Institutional Degrees to Open Association of Research, Society U.S.A.

Once you are designated as MARSC, you may send us a scanned copy of all of your credentials. OARS will verify, grade and certify them. This will be based on your academic records, quality of research papers published by you, and some more criteria.

It is mandatory to read all terms and conditions carefully.

Auxiliary Memberships

Institutional Fellow of Open Association of Research Society (USA)-OARS (USA)

Global Journals Incorporation (USA) is accredited by Open Association of Research Society, U.S.A (OARS) and in turn, affiliates research institutions as "Institutional Fellow of Open Association of Research Society" (IFOARS).

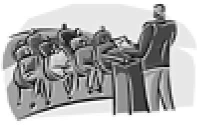
The "FARSC" is a dignified title which is accorded to a person's name viz. Dr. John E. Hall, Ph.D., FARSC or William Walldroff, M.S., FARSC.

The IFOARS institution is entitled to form a Board comprised of one Chairperson and three to five board members preferably from different streams. The Board will be recognized as "Institutional Board of Open Association of Research Society"-(IBOARS).

The Institute will be entitled to following benefits:

The IBOARS can initially review research papers of their institute and recommend them to publish with respective journal of Global Journals. It can also review the papers of other institutions after obtaining our consent. The second review will be done by peer reviewer of Global Journals Incorporation (USA) The Board is at liberty to appoint a peer reviewer with the approval of chairperson after consulting us.
The author fees of such paper may be waived off up to 40%.

The Global Journals Incorporation (USA) at its discretion can also refer double blind peer reviewed paper at their end to the board for the verification and to get recommendation for final stage of acceptance of publication.

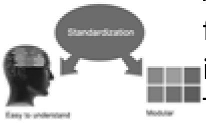


The IBOARS can organize symposium/seminar/conference in their country on behalf of Global Journals Incorporation (USA)-OARS (USA). The terms and conditions can be discussed separately.

The Board can also play vital role by exploring and giving valuable suggestions regarding the Standards of "Open Association of Research Society, U.S.A (OARS)" so that proper amendment can take place for the benefit of entire research community. We shall provide details of particular standard only on receipt of request from the Board.

The board members can also join us as Individual Fellow with 40% discount on total fees applicable to Individual Fellow. They will be entitled to avail all the benefits as declared. Please visit Individual Fellow-sub menu of GlobalJournals.org to have more relevant details.

We shall provide you intimation regarding launching of e-version of journal of your stream time to time. This may be utilized in your library for the enrichment of knowledge of your students as well as it can also be helpful for the concerned faculty members.



After nomination of your institution as “Institutional Fellow” and constantly functioning successfully for one year, we can consider giving recognition to your institute to function as Regional/Zonal office on our behalf.
The board can also take up the additional allied activities for betterment after our consultation.

The following entitlements are applicable to individual Fellows:

Open Association of Research Society, U.S.A (OARS) By-laws states that an individual Fellow may use the designations as applicable, or the corresponding initials. The Credentials of individual Fellow and Associate designations signify that the individual has gained knowledge of the fundamental concepts. One is magnanimous and proficient in an expertise course covering the professional code of conduct, and follows recognized standards of practice.

Open Association of Research Society (US)/ Global Journals Incorporation (USA), as described in Corporate Statements, are educational, research publishing and professional membership organizations. Achieving our individual Fellow or Associate status is based mainly on meeting stated educational research requirements.

Disbursement of 40% Royalty earned through Global Journals : Researcher = 50%, Peer Reviewer = 37.50%, Institution = 12.50% E.g. Out of 40%, the 20% benefit should be passed on to researcher, 15 % benefit towards remuneration should be given to a reviewer and remaining 5% is to be retained by the institution.

We shall provide print version of 12 issues of any three journals [as per your requirement] out of our 38 journals worth $ 2376 USD.

Other:

The individual Fellow and Associate designations accredited by Open Association of Research Society (US) credentials signify guarantees following achievements:

- The professional accredited with Fellow honor, is entitled to various benefits viz. name, fame, honor, regular flow of income, secured bright future, social status etc.

- In addition to above, if one is single author, then entitled to 40% discount on publishing research paper and can get 10%discount if one is co-author or main author among group of authors.
- The Fellow can organize symposium/seminar/conference on behalf of Global Journals Incorporation (USA) and he/she can also attend the same organized by other institutes on behalf of Global Journals.
- The Fellow can become member of Editorial Board Member after completing 3yrs.
- The Fellow can earn 60% of sales proceeds from the sale of reference/review books/literature/publishing of research paper.
- Fellow can also join as paid peer reviewer and earn 15% remuneration of author charges and can also get an opportunity to join as member of the Editorial Board of Global Journals Incorporation (USA)
- • This individual has learned the basic methods of applying those concepts and techniques to common challenging situations. This individual has further demonstrated an in–depth understanding of the application of suitable techniques to a particular area of research practice.

Note :

> "
>
> - In future, if the board feels the necessity to change any board member, the same can be done with the consent of the chairperson along with anyone board member without our approval.
> - In case, the chairperson needs to be replaced then consent of 2/3rd board members are required and they are also required to jointly pass the resolution copy of which should be sent to us. In such case, it will be compulsory to obtain our approval before replacement.
> - In case of "Difference of Opinion [if any]" among the Board members, our decision will be final and binding to everyone. "

Process of submission of Research Paper

The Area or field of specialization may or may not be of any category as mentioned in 'Scope of Journal' menu of the GlobalJournals.org website. There are 37 Research Journal categorized with Six parental Journals GJCST, GJMR, GJRE, GJMBR, GJSFR, GJHSS. For Authors should prefer the mentioned categories. There are three widely used systems UDC, DDC and LCC. The details are available as 'Knowledge Abstract' at Home page. The major advantage of this coding is that, the research work will be exposed to and shared with all over the world as we are being abstracted and indexed worldwide.

The paper should be in proper format. The format can be downloaded from first page of 'Author Guideline' Menu. The Author is expected to follow the general rules as mentioned in this menu. The paper should be written in MS-Word Format (*.DOC,*.DOCX).

The Author can submit the paper either online or offline. The authors should prefer online submission.Online Submission: There are three ways to submit your paper:

(A) (I) First, register yourself using top right corner of Home page then Login. If you are already registered, then login using your username and password.

(II) Choose corresponding Journal.

(III) Click 'Submit Manuscript'. Fill required information and Upload the paper.

(B) If you are using Internet Explorer, then Direct Submission through Homepage is also available.

(C) If these two are not convenient, and then email the paper directly to dean@globaljournals.org.

Offline Submission: Author can send the typed form of paper by Post. However, online submission should be preferred.

Preferred Author Guidelines

MANUSCRIPT STYLE INSTRUCTION (Must be strictly followed)

Page Size: 8.27" X 11'"

- Left Margin: 0.65
- Right Margin: 0.65
- Top Margin: 0.75
- Bottom Margin: 0.75
- Font type of all text should be Swis 721 Lt BT.
- Paper Title should be of Font Size 24 with one Column section.
- Author Name in Font Size of 11 with one column as of Title.
- Abstract Font size of 9 Bold, "Abstract" word in Italic Bold.
- Main Text: Font size 10 with justified two columns section
- Two Column with Equal Column with of 3.38 and Gaping of .2
- First Character must be three lines Drop capped.
- Paragraph before Spacing of 1 pt and After of 0 pt.
- Line Spacing of 1 pt
- Large Images must be in One Column
- Numbering of First Main Headings (Heading 1) must be in Roman Letters, Capital Letter, and Font Size of 10.
- Numbering of Second Main Headings (Heading 2) must be in Alphabets, Italic, and Font Size of 10.

You can use your own standard format also.
Author Guidelines:

1. General,

2. Ethical Guidelines,

3. Submission of Manuscripts,

4. Manuscript's Category,

5. Structure and Format of Manuscript,

6. After Acceptance.

1. GENERAL

Before submitting your research paper, one is advised to go through the details as mentioned in following heads. It will be beneficial, while peer reviewer justify your paper for publication.

Scope

The Global Journals Inc. (US) welcome the submission of original paper, review paper, survey article relevant to the all the streams of Philosophy and knowledge. The Global Journals Inc. (US) is parental platform for Global Journal of Computer Science and Technology, Researches in Engineering, Medical Research, Science Frontier Research, Human Social Science, Management, and Business organization. The choice of specific field can be done otherwise as following in Abstracting and Indexing Page on this Website. As the all Global

Journals Inc. (US) are being abstracted and indexed (in process) by most of the reputed organizations. Topics of only narrow interest will not be accepted unless they have wider potential or consequences.

2. ETHICAL GUIDELINES

Authors should follow the ethical guidelines as mentioned below for publication of research paper and research activities.

Papers are accepted on strict understanding that the material in whole or in part has not been, nor is being, considered for publication elsewhere. If the paper once accepted by Global Journals Inc. (US) and Editorial Board, will become the copyright of the Global Journals Inc. (US).

Authorship: The authors and coauthors should have active contribution to conception design, analysis and interpretation of findings. They should critically review the contents and drafting of the paper. All should approve the final version of the paper before submission

The Global Journals Inc. (US) follows the definition of authorship set up by the Global Academy of Research and Development. According to the Global Academy of R&D authorship, criteria must be based on:

1) Substantial contributions to conception and acquisition of data, analysis and interpretation of the findings.

2) Drafting the paper and revising it critically regarding important academic content.

3) Final approval of the version of the paper to be published.

All authors should have been credited according to their appropriate contribution in research activity and preparing paper. Contributors who do not match the criteria as authors may be mentioned under Acknowledgement.

Acknowledgements: Contributors to the research other than authors credited should be mentioned under acknowledgement. The specifications of the source of funding for the research if appropriate can be included. Suppliers of resources may be mentioned along with address.

Appeal of Decision: The Editorial Board's decision on publication of the paper is final and cannot be appealed elsewhere.

Permissions: It is the author's responsibility to have prior permission if all or parts of earlier published illustrations are used in this paper.

Please mention proper reference and appropriate acknowledgements wherever expected.

If all or parts of previously published illustrations are used, permission must be taken from the copyright holder concerned. It is the author's responsibility to take these in writing.

Approval for reproduction/modification of any information (including figures and tables) published elsewhere must be obtained by the authors/copyright holders before submission of the manuscript. Contributors (Authors) are responsible for any copyright fee involved.

3. SUBMISSION OF MANUSCRIPTS

Manuscripts should be uploaded via this online submission page. The online submission is most efficient method for submission of papers, as it enables rapid distribution of manuscripts and consequently speeds up the review procedure. It also enables authors to know the status of their own manuscripts by emailing us. Complete instructions for submitting a paper is available below.

Manuscript submission is a systematic procedure and little preparation is required beyond having all parts of your manuscript in a given format and a computer with an Internet connection and a Web browser. Full help and instructions are provided on-screen. As an author, you will be prompted for login and manuscript details as Field of Paper and then to upload your manuscript file(s) according to the instructions.

To avoid postal delays, all transaction is preferred by e-mail. A finished manuscript submission is confirmed by e-mail immediately and your paper enters the editorial process with no postal delays. When a conclusion is made about the publication of your paper by our Editorial Board, revisions can be submitted online with the same procedure, with an occasion to view and respond to all comments.

Complete support for both authors and co-author is provided.

4. MANUSCRIPT'S CATEGORY

Based on potential and nature, the manuscript can be categorized under the following heads:

Original research paper: Such papers are reports of high-level significant original research work.

Review papers: These are concise, significant but helpful and decisive topics for young researchers.

Research articles: These are handled with small investigation and applications.

Research letters: The letters are small and concise comments on previously published matters.

5. STRUCTURE AND FORMAT OF MANUSCRIPT

The recommended size of original research paper is less than seven thousand words, review papers fewer than seven thousands words also.Preparation of research paper or how to write research paper, are major hurdle, while writing manuscript. The research articles and research letters should be fewer than three thousand words, the structure original research paper; sometime review paper should be as follows:

Papers: These are reports of significant research (typically less than 7000 words equivalent, including tables, figures, references), and comprise:

(a)Title should be relevant and commensurate with the theme of the paper.

(b) A brief Summary, "Abstract" (less than 150 words) containing the major results and conclusions.

(c) Up to ten keywords, that precisely identifies the paper's subject, purpose, and focus.

(d) An Introduction, giving necessary background excluding subheadings; objectives must be clearly declared.

(e) Resources and techniques with sufficient complete experimental details (wherever possible by reference) to permit repetition; sources of information must be given and numerical methods must be specified by reference, unless non-standard.

(f) Results should be presented concisely, by well-designed tables and/or figures; the same data may not be used in both; suitable statistical data should be given. All data must be obtained with attention to numerical detail in the planning stage. As reproduced design has been recognized to be important to experiments for a considerable time, the Editor has decided that any paper that appears not to have adequate numerical treatments of the data will be returned un-refereed;

(g) Discussion should cover the implications and consequences, not just recapitulating the results; conclusions should be summarizing.

(h) Brief Acknowledgements.

(i) References in the proper form.

Authors should very cautiously consider the preparation of papers to ensure that they communicate efficiently. Papers are much more likely to be accepted, if they are cautiously designed and laid out, contain few or no errors, are summarizing, and be conventional to the approach and instructions. They will in addition, be published with much less delays than those that require much technical and editorial correction.

The Editorial Board reserves the right to make literary corrections and to make suggestions to improve briefness.

It is vital, that authors take care in submitting a manuscript that is written in simple language and adheres to published guidelines.

Format

Language: The language of publication is UK English. Authors, for whom English is a second language, must have their manuscript efficiently edited by an English-speaking person before submission to make sure that, the English is of high excellence. It is preferable, that manuscripts should be professionally edited.

Standard Usage, Abbreviations, and Units: Spelling and hyphenation should be conventional to The Concise Oxford English Dictionary. Statistics and measurements should at all times be given in figures, e.g. 16 min, except for when the number begins a sentence. When the number does not refer to a unit of measurement it should be spelt in full unless, it is 160 or greater.

Abbreviations supposed to be used carefully. The abbreviated name or expression is supposed to be cited in full at first usage, followed by the conventional abbreviation in parentheses.

Metric SI units are supposed to generally be used excluding where they conflict with current practice or are confusing. For illustration, 1.4 l rather than 1.4 × 10-3 m3, or 4 mm somewhat than 4 × 10-3 m. Chemical formula and solutions must identify the form used, e.g. anhydrous or hydrated, and the concentration must be in clearly defined units. Common species names should be followed by underlines at the first mention. For following use the generic name should be constricted to a single letter, if it is clear.

Structure

All manuscripts submitted to Global Journals Inc. (US), ought to include:

Title: The title page must carry an instructive title that reflects the content, a running title (less than 45 characters together with spaces), names of the authors and co-authors, and the place(s) wherever the work was carried out. The full postal address in addition with the e-mail address of related author must be given. Up to eleven keywords or very brief phrases have to be given to help data retrieval, mining and indexing.

Abstract, used in Original Papers and Reviews:

Optimizing Abstract for Search Engines

Many researchers searching for information online will use search engines such as Google, Yahoo or similar. By optimizing your paper for search engines, you will amplify the chance of someone finding it. This in turn will make it more likely to be viewed and/or cited in a further work. Global Journals Inc. (US) have compiled these guidelines to facilitate you to maximize the web-friendliness of the most public part of your paper.

Key Words

A major linchpin in research work for the writing research paper is the keyword search, which one will employ to find both library and Internet resources.

One must be persistent and creative in using keywords. An effective keyword search requires a strategy and planning a list of possible keywords and phrases to try.

Search engines for most searches, use Boolean searching, which is somewhat different from Internet searches. The Boolean search uses "operators," words (and, or, not, and near) that enable you to expand or narrow your affords. Tips for research paper while preparing research paper are very helpful guideline of research paper.

Choice of key words is first tool of tips to write research paper. Research paper writing is an art.A few tips for deciding as strategically as possible about keyword search:

- One should start brainstorming lists of possible keywords before even begin searching. Think about the most important concepts related to research work. Ask, "What words would a source have to include to be truly valuable in research paper?" Then consider synonyms for the important words.
- It may take the discovery of only one relevant paper to let steer in the right keyword direction because in most databases, the keywords under which a research paper is abstracted are listed with the paper.
- One should avoid outdated words.

Keywords are the key that opens a door to research work sources. Keyword searching is an art in which researcher's skills are bound to improve with experience and time.

Numerical Methods: Numerical methods used should be clear and, where appropriate, supported by references.

Acknowledgements: Please make these as concise as possible.

References

References follow the Harvard scheme of referencing. References in the text should cite the authors' names followed by the time of their publication, unless there are three or more authors when simply the first author's name is quoted followed by et al. unpublished work has to only be cited where necessary, and only in the text. Copies of references in press in other journals have to be supplied with submitted typescripts. It is necessary that all citations and references be carefully checked before submission, as mistakes or omissions will cause delays.

References to information on the World Wide Web can be given, but only if the information is available without charge to readers on an official site. Wikipedia and Similar websites are not allowed where anyone can change the information. Authors will be asked to make available electronic copies of the cited information for inclusion on the Global Journals Inc. (US) homepage at the judgment of the Editorial Board.

The Editorial Board and Global Journals Inc. (US) recommend that, citation of online-published papers and other material should be done via a DOI (digital object identifier). If an author cites anything, which does not have a DOI, they run the risk of the cited material not being noticeable.

The Editorial Board and Global Journals Inc. (US) recommend the use of a tool such as Reference Manager for reference management and formatting.

Tables, Figures and Figure Legends

Tables: Tables should be few in number, cautiously designed, uncrowned, and include only essential data. Each must have an Arabic number, e.g. Table 4, a self-explanatory caption and be on a separate sheet. Vertical lines should not be used.

Figures: Figures are supposed to be submitted as separate files. Always take in a citation in the text for each figure using Arabic numbers, e.g. Fig. 4. Artwork must be submitted online in electronic form by e-mailing them.

Preparation of Electronic Figures for Publication

Even though low quality images are sufficient for review purposes, print publication requires high quality images to prevent the final product being blurred or fuzzy. Submit (or e-mail) EPS (line art) or TIFF (halftone/photographs) files only. MS PowerPoint and Word Graphics are unsuitable for printed pictures. Do not use pixel-oriented software. Scans (TIFF only) should have a resolution of at least 350 dpi (halftone) or 700 to 1100 dpi (line drawings) in relation to the imitation size. Please give the data for figures in black and white or submit a Color Work Agreement Form. EPS files must be saved with fonts embedded (and with a TIFF preview, if possible).

For scanned images, the scanning resolution (at final image size) ought to be as follows to ensure good reproduction: line art: >650 dpi; halftones (including gel photographs) : >350 dpi; figures containing both halftone and line images: >650 dpi.

Color Charges: It is the rule of the Global Journals Inc. (US) for authors to pay the full cost for the reproduction of their color artwork. Hence, please note that, if there is color artwork in your manuscript when it is accepted for publication, we would require you to complete and return a color work agreement form before your paper can be published.

Figure Legends: Self-explanatory legends of all figures should be incorporated separately under the heading 'Legends to Figures'. In the full-text online edition of the journal, figure legends may possibly be truncated in abbreviated links to the full screen version. Therefore, the first 100 characters of any legend should notify the reader, about the key aspects of the figure.

6. AFTER ACCEPTANCE

Upon approval of a paper for publication, the manuscript will be forwarded to the dean, who is responsible for the publication of the Global Journals Inc. (US).

6.1 Proof Corrections

The corresponding author will receive an e-mail alert containing a link to a website or will be attached. A working e-mail address must therefore be provided for the related author.

Acrobat Reader will be required in order to read this file. This software can be downloaded

(Free of charge) from the following website:

www.adobe.com/products/acrobat/readstep2.html. This will facilitate the file to be opened, read on screen, and printed out in order for any corrections to be added. Further instructions will be sent with the proof.

Proofs must be returned to the dean at dean@globaljournals.org within three days of receipt.

As changes to proofs are costly, we inquire that you only correct typesetting errors. All illustrations are retained by the publisher. Please note that the authors are responsible for all statements made in their work, including changes made by the copy editor.

6.2 Early View of Global Journals Inc. (US) (Publication Prior to Print)

The Global Journals Inc. (US) are enclosed by our publishing's Early View service. Early View articles are complete full-text articles sent in advance of their publication. Early View articles are absolute and final. They have been completely reviewed, revised and edited for publication, and the authors' final corrections have been incorporated. Because they are in final form, no changes can be made after sending them. The nature of Early View articles means that they do not yet have volume, issue or page numbers, so Early View articles cannot be cited in the conventional way.

6.3 Author Services

Online production tracking is available for your article through Author Services. Author Services enables authors to track their article - once it has been accepted - through the production process to publication online and in print. Authors can check the status of their articles online and choose to receive automated e-mails at key stages of production. The authors will receive an e-mail with a unique link that enables them to register and have their article automatically added to the system. Please ensure that a complete e-mail address is provided when submitting the manuscript.

6.4 Author Material Archive Policy

Please note that if not specifically requested, publisher will dispose off hardcopy & electronic information submitted, after the two months of publication. If you require the return of any information submitted, please inform the Editorial Board or dean as soon as possible.

6.5 Offprint and Extra Copies

A PDF offprint of the online-published article will be provided free of charge to the related author, and may be distributed according to the Publisher's terms and conditions. Additional paper offprint may be ordered by emailing us at: editor@globaljournals.org .

You must strictly follow above Author Guidelines before submitting your paper or else we will not at all be responsible for any corrections in future in any of the way.

Before start writing a good quality Computer Science Research Paper, let us first understand what is Computer Science Research Paper? So, Computer Science Research Paper is the paper which is written by professionals or scientists who are associated to Computer Science and Information Technology, or doing research study in these areas. If you are novel to this field then you can consult about this field from your supervisor or guide.

TECHNIQUES FOR WRITING A GOOD QUALITY RESEARCH PAPER:

1. Choosing the topic: In most cases, the topic is searched by the interest of author but it can be also suggested by the guides. You can have several topics and then you can judge that in which topic or subject you are finding yourself most comfortable. This can be done by asking several questions to yourself, like Will I be able to carry our search in this area? Will I find all necessary recourses to accomplish the search? Will I be able to find all information in this field area? If the answer of these types of questions will be "Yes" then you can choose that topic. In most of the cases, you may have to conduct the surveys and have to visit several places because this field is related to Computer Science and Information Technology. Also, you may have to do a lot of work to find all rise and falls regarding the various data of that subject. Sometimes, detailed information plays a vital role, instead of short information.

2. Evaluators are human: First thing to remember that evaluators are also human being. They are not only meant for rejecting a paper. They are here to evaluate your paper. So, present your Best.

3. Think Like Evaluators: If you are in a confusion or getting demotivated that your paper will be accepted by evaluators or not, then think and try to evaluate your paper like an Evaluator. Try to understand that what an evaluator wants in your research paper and automatically you will have your answer.

4. Make blueprints of paper: The outline is the plan or framework that will help you to arrange your thoughts. It will make your paper logical. But remember that all points of your outline must be related to the topic you have chosen.

5. Ask your Guides: If you are having any difficulty in your research, then do not hesitate to share your difficulty to your guide (if you have any). They will surely help you out and resolve your doubts. If you can't clarify what exactly you require for your work then ask the supervisor to help you with the alternative. He might also provide you the list of essential readings.

6. Use of computer is recommended: As you are doing research in the field of Computer Science, then this point is quite obvious.

7. Use right software: Always use good quality software packages. If you are not capable to judge good software then you can lose quality of your paper unknowingly. There are various software programs available to help you, which you can get through Internet.

8. Use the Internet for help: An excellent start for your paper can be by using the Google. It is an excellent search engine, where you can have your doubts resolved. You may also read some answers for the frequent question how to write my research paper or find model research paper. From the internet library you can download books. If you have all required books make important reading selecting and analyzing the specified information. Then put together research paper sketch out.

9. Use and get big pictures: Always use encyclopedias, Wikipedia to get pictures so that you can go into the depth.

10. Bookmarks are useful: When you read any book or magazine, you generally use bookmarks, right! It is a good habit, which helps to not to lose your continuity. You should always use bookmarks while searching on Internet also, which will make your search easier.

11. Revise what you wrote: When you write anything, always read it, summarize it and then finalize it.

12. Make all efforts: Make all efforts to mention what you are going to write in your paper. That means always have a good start. Try to mention everything in introduction, that what is the need of a particular research paper. Polish your work by good skill of writing and always give an evaluator, what he wants.

13. Have backups: When you are going to do any important thing like making research paper, you should always have backup copies of it either in your computer or in paper. This will help you to not to lose any of your important.

14. Produce good diagrams of your own: Always try to include good charts or diagrams in your paper to improve quality. Using several and unnecessary diagrams will degrade the quality of your paper by creating "hotchpotch." So always, try to make and include those diagrams, which are made by your own to improve readability and understandability of your paper.

15. Use of direct quotes: When you do research relevant to literature, history or current affairs then use of quotes become essential but if study is relevant to science then use of quotes is not preferable.

16. Use proper verb tense: Use proper verb tenses in your paper. Use past tense, to present those events that happened. Use present tense to indicate events that are going on. Use future tense to indicate future happening events. Use of improper and wrong tenses will confuse the evaluator. Avoid the sentences that are incomplete.

17. Never use online paper: If you are getting any paper on Internet, then never use it as your research paper because it might be possible that evaluator has already seen it or maybe it is outdated version.

18. Pick a good study spot: To do your research studies always try to pick a spot, which is quiet. Every spot is not for studies. Spot that suits you choose it and proceed further.

19. Know what you know: Always try to know, what you know by making objectives. Else, you will be confused and cannot achieve your target.

20. Use good quality grammar: Always use a good quality grammar and use words that will throw positive impact on evaluator. Use of good quality grammar does not mean to use tough words, that for each word the evaluator has to go through dictionary. Do not start sentence with a conjunction. Do not fragment sentences. Eliminate one-word sentences. Ignore passive voice. Do not ever use a big word when a diminutive one would suffice. Verbs have to be in agreement with their subjects. Prepositions are not expressions to finish sentences with. It is incorrect to ever divide an infinitive. Avoid clichés like the disease. Also, always shun irritating alliteration. Use language that is simple and straight forward. put together a neat summary.

21. Arrangement of information: Each section of the main body should start with an opening sentence and there should be a changeover at the end of the section. Give only valid and powerful arguments to your topic. You may also maintain your arguments with records.

22. Never start in last minute: Always start at right time and give enough time to research work. Leaving everything to the last minute will degrade your paper and spoil your work.

23. Multitasking in research is not good: Doing several things at the same time proves bad habit in case of research activity. Research is an area, where everything has a particular time slot. Divide your research work in parts and do particular part in particular time slot.

24. Never copy others' work: Never copy others' work and give it your name because if evaluator has seen it anywhere you will be in trouble.

25. Take proper rest and food: No matter how many hours you spend for your research activity, if you are not taking care of your health then all your efforts will be in vain. For a quality research, study is must, and this can be done by taking proper rest and food.

26. Go for seminars: Attend seminars if the topic is relevant to your research area. Utilize all your resources.

27. Refresh your mind after intervals: Try to give rest to your mind by listening to soft music or by sleeping in intervals. This will also improve your memory.

28. Make colleagues: Always try to make colleagues. No matter how sharper or intelligent you are, if you make colleagues you can have several ideas, which will be helpful for your research.

29. Think technically: Always think technically. If anything happens, then search its reasons, its benefits, and demerits.

30. Think and then print: When you will go to print your paper, notice that tables are not be split, headings are not detached from their descriptions, and page sequence is maintained.

31. Adding unnecessary information: Do not add unnecessary information, like, I have used MS Excel to draw graph. Do not add irrelevant and inappropriate material. These all will create superfluous. Foreign terminology and phrases are not apropos. One should NEVER take a broad view. Analogy in script is like feathers on a snake. Not at all use a large word when a very small one would be sufficient. Use words properly, regardless of how others use them. Remove quotations. Puns are for kids, not grunt readers. Amplification is a billion times of inferior quality than sarcasm.

32. Never oversimplify everything: To add material in your research paper, never go for oversimplification. This will definitely irritate the evaluator. Be more or less specific. Also too, by no means, ever use rhythmic redundancies. Contractions aren't essential and shouldn't be there used. Comparisons are as terrible as clichés. Give up ampersands and abbreviations, and so on. Remove commas, that are, not necessary. Parenthetical words however should be together with this in commas. Understatement is all the time the complete best way to put onward earth-shaking thoughts. Give a detailed literary review.

33. Report concluded results: Use concluded results. From raw data, filter the results and then conclude your studies based on measurements and observations taken. Significant figures and appropriate number of decimal places should be used. Parenthetical remarks are prohibitive. Proofread carefully at final stage. In the end give outline to your arguments. Spot out perspectives of further study of this subject. Justify your conclusion by at the bottom of them with sufficient justifications and examples.

34. After conclusion: Once you have concluded your research, the next most important step is to present your findings. Presentation is extremely important as it is the definite medium though which your research is going to be in print to the rest of the crowd. Care should be taken to categorize your thoughts well and present them in a logical and neat manner. A good quality research paper format is essential because it serves to highlight your research paper and bring to light all necessary aspects in your research.

Informal Guidelines of Research Paper Writing

Key points to remember:

- Submit all work in its final form.
- Write your paper in the form, which is presented in the guidelines using the template.
- Please note the criterion for grading the final paper by peer-reviewers.

Final Points:

A purpose of organizing a research paper is to let people to interpret your effort selectively. The journal requires the following sections, submitted in the order listed, each section to start on a new page.

The introduction will be compiled from reference matter and will reflect the design processes or outline of basis that direct you to make study. As you will carry out the process of study, the method and process section will be constructed as like that. The result segment will show related statistics in nearly sequential order and will direct the reviewers next to the similar intellectual paths throughout the data that you took to carry out your study. The discussion section will provide understanding of the data and projections as to the implication of the results. The use of good quality references all through the paper will give the effort trustworthiness by representing an alertness of prior workings.

Writing a research paper is not an easy job no matter how trouble-free the actual research or concept. Practice, excellent preparation, and controlled record keeping are the only means to make straightforward the progression.

General style:

Specific editorial column necessities for compliance of a manuscript will always take over from directions in these general guidelines.

To make a paper clear

· Adhere to recommended page limits

Mistakes to evade

- Insertion a title at the foot of a page with the subsequent text on the next page
- Separating a table/chart or figure - impound each figure/table to a single page
- Submitting a manuscript with pages out of sequence

In every sections of your document

· Use standard writing style including articles ("a", "the," etc.)

· Keep on paying attention on the research topic of the paper

· Use paragraphs to split each significant point (excluding for the abstract)

· Align the primary line of each section

· Present your points in sound order

· Use present tense to report well accepted

· Use past tense to describe specific results

· Shun familiar wording, don't address the reviewer directly, and don't use slang, slang language, or superlatives

· Shun use of extra pictures - include only those figures essential to presenting results

Title Page:

Choose a revealing title. It should be short. It should not have non-standard acronyms or abbreviations. It should not exceed two printed lines. It should include the name(s) and address (es) of all authors.

Abstract:

The summary should be two hundred words or less. It should briefly and clearly explain the key findings reported in the manuscript--must have precise statistics. It should not have abnormal acronyms or abbreviations. It should be logical in itself. Shun citing references at this point.

An abstract is a brief distinct paragraph summary of finished work or work in development. In a minute or less a reviewer can be taught the foundation behind the study, common approach to the problem, relevant results, and significant conclusions or new questions.

Write your summary when your paper is completed because how can you write the summary of anything which is not yet written? Wealth of terminology is very essential in abstract. Yet, use comprehensive sentences and do not let go readability for briefness. You can maintain it succinct by phrasing sentences so that they provide more than lone rationale. The author can at this moment go straight to shortening the outcome. Sum up the study, with the subsequent elements in any summary. Try to maintain the initial two items to no more than one ruling each.

- Reason of the study - theory, overall issue, purpose
- Fundamental goal
- To the point depiction of the research
- Consequences, including definite statistics - if the consequences are quantitative in nature, account quantitative data; results of any numerical analysis should be reported
- Significant conclusions or questions that track from the research(es)

Approach:

- Single section, and succinct
- As a outline of job done, it is always written in past tense
- A conceptual should situate on its own, and not submit to any other part of the paper such as a form or table
- Center on shortening results - bound background information to a verdict or two, if completely necessary
- What you account in an conceptual must be regular with what you reported in the manuscript
- Exact spelling, clearness of sentences and phrases, and appropriate reporting of quantities (proper units, important statistics) are just as significant in an abstract as they are anywhere else

Introduction:

The **Introduction** should "introduce" the manuscript. The reviewer should be presented with sufficient background information to be capable to comprehend and calculate the purpose of your study without having to submit to other works. The basis for the study should be offered. Give most important references but shun difficult to make a comprehensive appraisal of the topic. In the introduction, describe the problem visibly. If the problem is not acknowledged in a logical, reasonable way, the reviewer will have no attention in your result. Speak in common terms about techniques used to explain the problem, if needed, but do not present any particulars about the protocols here. Following approach can create a valuable beginning:

- Explain the value (significance) of the study
- Shield the model - why did you employ this particular system or method? What is its compensation? You strength remark on its appropriateness from a abstract point of vision as well as point out sensible reasons for using it.
- Present a justification. Status your particular theory (es) or aim(s), and describe the logic that led you to choose them.
- Very for a short time explain the tentative propose and how it skilled the declared objectives.

Approach:

- Use past tense except for when referring to recognized facts. After all, the manuscript will be submitted after the entire job is done.
- Sort out your thoughts; manufacture one key point with every section. If you make the four points listed above, you will need a least of four paragraphs.

- Present surroundings information only as desirable in order hold up a situation. The reviewer does not desire to read the whole thing you know about a topic.
- Shape the theory/purpose specifically - do not take a broad view.
- As always, give awareness to spelling, simplicity and correctness of sentences and phrases.

Procedures (Methods and Materials):

This part is supposed to be the easiest to carve if you have good skills. A sound written Procedures segment allows a capable scientist to replacement your results. Present precise information about your supplies. The suppliers and clarity of reagents can be helpful bits of information. Present methods in sequential order but linked methodologies can be grouped as a segment. Be concise when relating the protocols. Attempt for the least amount of information that would permit another capable scientist to spare your outcome but be cautious that vital information is integrated. The use of subheadings is suggested and ought to be synchronized with the results section. When a technique is used that has been well described in another object, mention the specific item describing a way but draw the basic principle while stating the situation. The purpose is to text all particular resources and broad procedures, so that another person may use some or all of the methods in one more study or referee the scientific value of your work. It is not to be a step by step report of the whole thing you did, nor is a methods section a set of orders.

Materials:

- Explain materials individually only if the study is so complex that it saves liberty this way.
- Embrace particular materials, and any tools or provisions that are not frequently found in laboratories.
- Do not take in frequently found.
- If use of a definite type of tools.
- Materials may be reported in a part section or else they may be recognized along with your measures.

Methods:

- Report the method (not particulars of each process that engaged the same methodology)
- Describe the method entirely
- To be succinct, present methods under headings dedicated to specific dealings or groups of measures
- Simplify - details how procedures were completed not how they were exclusively performed on a particular day.
- If well known procedures were used, account the procedure by name, possibly with reference, and that's all.

Approach:

- It is embarrassed or not possible to use vigorous voice when documenting methods with no using first person, which would focus the reviewer's interest on the researcher rather than the job. As a result when script up the methods most authors use third person passive voice.
- Use standard style in this and in every other part of the paper - avoid familiar lists, and use full sentences.

What to keep away from

- Resources and methods are not a set of information.
- Skip all descriptive information and surroundings - save it for the argument.
- Leave out information that is immaterial to a third party.

Results:

The principle of a results segment is to present and demonstrate your conclusion. Create this part a entirely objective details of the outcome, and save all understanding for the discussion.

The page length of this segment is set by the sum and types of data to be reported. Carry on to be to the point, by means of statistics and tables, if suitable, to present consequences most efficiently.You must obviously differentiate material that would usually be incorporated in a study editorial from any unprocessed data or additional appendix matter that would not be available. In fact, such matter should not be submitted at all except requested by the instructor.

Content

- Sum up your conclusion in text and demonstrate them, if suitable, with figures and tables.
- In manuscript, explain each of your consequences, point the reader to remarks that are most appropriate.
- Present a background, such as by describing the question that was addressed by creation an exacting study.
- Explain results of control experiments and comprise remarks that are not accessible in a prescribed figure or table, if appropriate.
- Examine your data, then prepare the analyzed (transformed) data in the form of a figure (graph), table, or in manuscript form.

What to stay away from

- Do not discuss or infer your outcome, report surroundings information, or try to explain anything.
- Not at all, take in raw data or intermediate calculations in a research manuscript.
- Do not present the similar data more than once.
- Manuscript should complement any figures or tables, not duplicate the identical information.
- Never confuse figures with tables - there is a difference.

Approach

- As forever, use past tense when you submit to your results, and put the whole thing in a reasonable order.
- Put figures and tables, appropriately numbered, in order at the end of the report
- If you desire, you may place your figures and tables properly within the text of your results part.

Figures and tables

- If you put figures and tables at the end of the details, make certain that they are visibly distinguished from any attach appendix materials, such as raw facts
- Despite of position, each figure must be numbered one after the other and complete with subtitle
- In spite of position, each table must be titled, numbered one after the other and complete with heading
- All figure and table must be adequately complete that it could situate on its own, divide from text

Discussion:

The Discussion is expected the trickiest segment to write and describe. A lot of papers submitted for journal are discarded based on problems with the Discussion. There is no head of state for how long a argument should be. Position your understanding of the outcome visibly to lead the reviewer through your conclusions, and then finish the paper with a summing up of the implication of the study. The purpose here is to offer an understanding of your results and hold up for all of your conclusions, using facts from your research and generally accepted information, if suitable. The implication of result should be visibly described. Infer your data in the conversation in suitable depth. This means that when you clarify an observable fact you must explain mechanisms that may account for the observation. If your results vary from your prospect, make clear why that may have happened. If your results agree, then explain the theory that the proof supported. It is never suitable to just state that the data approved with prospect, and let it drop at that.

- Make a decision if each premise is supported, discarded, or if you cannot make a conclusion with assurance. Do not just dismiss a study or part of a study as "uncertain."
- Research papers are not acknowledged if the work is imperfect. Draw what conclusions you can based upon the results that you have, and take care of the study as a finished work
- You may propose future guidelines, such as how the experiment might be personalized to accomplish a new idea.
- Give details all of your remarks as much as possible, focus on mechanisms.
- Make a decision if the tentative design sufficiently addressed the theory, and whether or not it was correctly restricted.
- Try to present substitute explanations if sensible alternatives be present.
- One research will not counter an overall question, so maintain the large picture in mind, where do you go next? The best studies unlock new avenues of study. What questions remain?
- Recommendations for detailed papers will offer supplementary suggestions.

Approach:

- When you refer to information, differentiate data generated by your own studies from available information
- Submit to work done by specific persons (including you) in past tense.
- Submit to generally acknowledged facts and main beliefs in present tense.

The Administration Rules

Please carefully note down following rules and regulation before submitting your Research Paper to Global Journals Inc. (US):

Segment Draft and Final Research Paper: You have to strictly follow the template of research paper. If it is not done your paper may get rejected.

- The **major constraint** is that you must independently make all content, tables, graphs, and facts that are offered in the paper. You must write each part of the paper wholly on your own. The Peer-reviewers need to identify your own perceptive of the concepts in your own terms. NEVER extract straight from any foundation, and never rephrase someone else's analysis.

- Do not give permission to anyone else to "PROOFREAD" your manuscript.

- Methods to avoid Plagiarism is applied by us on every paper, if found guilty, you will be blacklisted by all of our collaborated research groups, your institution will be informed for this and strict legal actions will be taken immediately.)
- To guard yourself and others from possible illegal use please do not permit anyone right to use to your paper and files.

Criterion for Grading a Research Paper (Compilation)
by Global Journals Inc. (US)

Please note that following table is only a Grading of "Paper Compilation" and not on "Performed/Stated Research" whose grading solely depends on Individual Assigned Peer Reviewer and Editorial Board Member. These can be available only on request and after decision of Paper. This report will be the property of Global Journals Inc. (US).

Topics	Grades		
	A-B	C-D	E-F
Abstract	Clear and concise with appropriate content, Correct format. 200 words or below	Unclear summary and no specific data, Incorrect form Above 200 words	No specific data with ambiguous information Above 250 words
Introduction	Containing all background details with clear goal and appropriate details, flow specification, no grammar and spelling mistake, well organized sentence and paragraph, reference cited	Unclear and confusing data, appropriate format, grammar and spelling errors with unorganized matter	Out of place depth and content, hazy format
Methods and Procedures	Clear and to the point with well arranged paragraph, precision and accuracy of facts and figures, well organized subheads	Difficult to comprehend with embarrassed text, too much explanation but completed	Incorrect and unorganized structure with hazy meaning
Result	Well organized, Clear and specific, Correct units with precision, correct data, well structuring of paragraph, no grammar and spelling mistake	Complete and embarrassed text, difficult to comprehend	Irregular format with wrong facts and figures
Discussion	Well organized, meaningful specification, sound conclusion, logical and concise explanation, highly structured paragraph reference cited	Wordy, unclear conclusion, spurious	Conclusion is not cited, unorganized, difficult to comprehend
References	Complete and correct format, well organized	Beside the point, Incomplete	Wrong format and structuring

INDEX

save our planet

Global Journal of Computer Science and Technology

Visit us on the Web at www.GlobalJournals.org | www.ComputerResearch.org
or email us at helpdesk@globaljournals.org

ISSN 9754350