

Effective Approaches of Classification Algorithms for Text Mining Applications

U.Latha¹, T.Velmurugan²

¹Research Scholar, ²Associate Professor

PG and Research Department of Computer Science, D. G. Vaishnav College, Arumbakkam, Chennai, India

E-Mail: ¹dgvclatha@gmail.com, ²velmurugan_dgvc@yahoo.co.in

Abstract: The large amount of data stored in unstructured texts cannot simply be used for further processing by computers, which typically handle text as simple sequences of character strings. Therefore, specific (pre-) processing methods and algorithms are required in order to retrieve useful information via text. Text mining refers generally to the process of retrieving information and knowledge from formless text. This research work analyses about the use of classification algorithms and their uses to predict the applications of text mining. The purpose of

this work is to present an analysis of recent publications concerning with text mining using classification algorithm in particular. This survey finds out some of the best suitable algorithms for text mining analyses suggested by the various researchers in their research work.

Keywords: Classification Algorithms, Text Mining, CART Algorithm, ID3 Algorithm, C4.5 Algorithm

I. INTRODUCTION

Data Mining (DM) techniques are very useful to find the behavior of text in various domains, particularly in World Wide Web and some other repositories according to its usages. DM is also defined as Knowledge Discovery in Databases (KDD). The techniques in DM have used majorly preprocess analysis and pattern generation. DM has got more and more matures as a field of major research in information technology, computer science and got widely applied in several other fields. DM can be implemented on various types of databases and information repositories, but the kind of patterns to be found are specified by several data mining functionalities like class and concept description, association, correlation analysis, classification, prediction, cluster analysis etc. It can be performed on various types of clustering and classifications techniques. Classification is one of the most important subroutines in machine learning and data mining tasks. This research work mainly focusses on the text mining applications using classification algorithms.

These steps have to be performed iteratively and several steps usually require interactive feedback from a user [2]. As defined by the Cross Industry Standard Process for Data Mining (Crisp DM) model [cri99] the main steps are: (I) Business understanding, (II) Data understanding, (III) Data preparation, (IV) Modelling, (V) Evaluation, (VI) Deployment. Besides the initial problem of analyzing and understanding the overall task (first two steps) one of the most time consuming steps is data preparation. This is especially of interest for text mining which needs special preprocessing methods to convert textual data into a format which is suitable for data mining algorithms. The

application of data mining algorithms in the modelling step, the evaluation of the obtained model and the deployment of the application (if necessary) are closing the process cycle. Here the modelling step is of main interest as text mining frequently requires the development of new or the adaptation of existing algorithms.

Classification is a supervised Machine Learning technique which assigns labels or classes to not the same objects or groups. Classification is a two-step process First step is model construction which is defined as the analysis of the training records of a database. Second step is model usage to the constructed model is used for classification. The classification accuracy is estimated by the percentage of test samples or records that are correctly classified. In the Classification has been successfully applied to a wide range of application areas, such as scientific experiments, medical diagnosis, weather prediction, credit approval, customer segmentation, text mining, web mining, target marketing and fraud detection.

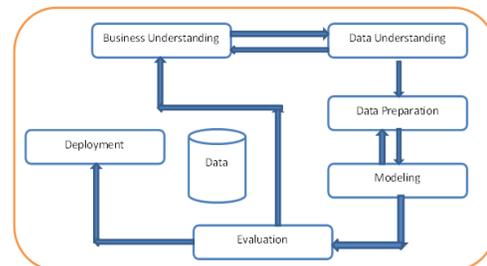


Figure 1: Phases of Crisp DM

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing domains. C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan. C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is often referred to as a statistical classifier. It became quite popular after ranking #1 in the Top 10 Algorithms in Data Mining pre-eminent paper published by Springer LNCS in 2008.

Decision trees used in data mining are of two main types: Classification tree analysis is when the predicted outcome is the class to which the data belongs. Regression tree analysis is when the predicted outcome can be considered a real number (e.g. the price of a house, or a patient's length of stay in a hospital). The term Classification And Regression Tree

(CART) analysis is an umbrella term used to refer to both of the above procedures, first introduced by Breiman et al. Trees used for regression and trees used for classification have some similarities - but also some differences, such as the procedure used to determine where to split.

The organization of this paper is structured as follows. Section II discusses about the applications classification algorithms in other domains. The text mining applications without using classification algorithms are explored in section III. Section IV elaborated about the text mining applications using classification algorithms. Finally, section V concludes the research work.

II. APPLICATIONS OF CLASSIFICATION ALGORITHM

The applications of classification algorithm covers basic concept of the data mining and the applications. A Classification algorithm is a set of rules that have been in detail written for execution on two or more processors. The data mining tool and its techniques are highlighted in text mining. Data mining is a great and a different field having various techniques in extracting text to analyses the recent real world problems. It converts the raw data into useful information in various research areas of extracting text. There are various major data mining techniques that have been developed and used in data mining projects recently for knowledge discovery from database.

Ian H. Witten et al. discussed about Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations in their research work [8]. They told about the usage of WEKA(Waikato Environment for Knowledge Analysis) tool in the field of data mining for classification algorithm, association rules and also clustering algorithms. They also analyzed that WEKA is the greatest tool of data mining which is implemented by JAVA. They finally conclude with the analysis of advanced topics such as visual machine learning, text mining, and Web mining. Survey analysis carried out by Brijain R Pat et al. in [4]. In this paper they found the various classifications algorithm of the decision tree ID3, C4.5, C5.0 algorithms to identify the classifying speed, strong learning ability and simple construction with the classification methods to find the classifying problems. When they use these algorithms to classify, there will be problems in choosing attribute which have more values, and overlooking attributes which have less value.

A research paper by Wenmin Li et al. discussed about a classification based on multiple class-association rules [29]. In their study, they proposed a new associative classification method, CMAR, i.e., Classification based on Multiple Association Rules. They Compared 3 classification algorithms such as C4.5, CBA and CMAR with 26 databases in UCI machine learning repository. They got the average accuracy of 84.69% in CBA, 83.34% in C4.5 and 85.22% in CMAR. They conclude that it is highly effective and better average accuracy in comparison with CBA and C4.5. A research work titled as Data Preparation for Mining World Wide Web Browsing Patterns is carried out in [20] by Robert cooley et al. This paper has presented the preprocessing details about Web Usage

Mining, they analyzed about the preprocessing tasks for WWW server access logs using classification algorithms and association rules. This paper has experimental results on synthetic data and on real world industrial data and the reference length is performed consistently. SLIQ: A Fast Scalable Classifier for Data Mining [12] explored by Manish Mehta et al. They tell that the classification is an important problem in data mining. They explored a new classification algorithm called SLIQ is a decision tree classifier. They compared CART, C4 and newly proposed algorithm SLIQ. They conclude that it produces accurate decision tree than the C4 and magnitude faster than CART. S. B. Kotsiantis Carried out a research work by a review paper titled as "Supervised Machine Learning: A Review of Classification Techniques" [21]. This paper tells about the newly proposed best-known supervised techniques known as Support Vector Machines (SVMs) from the relative detail; they conclude that the output coding approach, to reduce a multi-class problem to a set of multiple binary classification problems.

Another work carried out by Rafael S. Parpinelli, Heitor S. Lopes and Alex A. Freitas in [18]. This paper has explored an algorithm for rule discovery called Ant-Miner. They explored a new algorithm called Ant-Miner is to get classification rules in data sets. They have compared the performance of Ant-Miner and the CN2 algorithm in six public domain data sets. The experimental result tells that the predictive accuracy is better in four data sets regarding to Ant-Miner, one in CN2 and same in both. Therefore, it is comparatively better predictive accuracy in Ant-Miner than CN2. State-of-the-art in Privacy Preserving Data Mining is done by Vassilios S. Verykios et al. in [26].

They analyzed about the Heuristic Based Techniques, Cryptography Based Techniques and Reconstruction Based Techniques in privacy preserving data mining and they conclude that they have reviewing that area, the data mining algorithms have some limits according to the privacy issues can be effectively considered. Spatial Data Mining: A Database Approach [13] explored by Martin Ester et al. The aim of this paper is to define a set of basic operations for KDD in SDBS. They analyzed spatial association rules, clustering, trend detection and classifications. They explored the neighborhood graphs and paths for small set of operations and they conclude that neighborhood indices are introduced to materialize selected neighborhood graphs in order to processing speed up of the proposed operations.

A research work titled as Integrating Classification and Association Rule Mining is carried out in [3] by Bing Liu et al. This paper proposes a new way to combine the classification and association rule mining. An algorithm is explored to do all class association rules (CARs) and to construct an accurate classifier. This way is not only for the classifiers and also to solve the number of problems in the classification systems. A Data Mining Framework for Building Intrusion Detection Models carried out by Wenke Lee et al. [28]. In this paper, they analyzed about the construction of intrusion detection models. This research work tells that the observed behavior in the data and to apply data mining programs to review data to compute misuse and anomaly detection models. To facilitate adaptability and extensibility, they propose the use of meta-learning by means to build a

combined model that incorporate evidence from multiple base models.

III. TEXT MINING EXTRACTION IN DATA MINING

Text mining, also referred to as *text data mining*, roughly equivalent to text analytics, refers to the process of deriving high quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (*i.e.*, learning relations between named entities).

A research work titled as Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data is carried out by Jaideep Srivastava et al. in [9]. This paper has tells the up-to-date survey of Web Usage mining and it is consists of three parts which is preprocessing, pattern discovery and pattern analysis. This analysis serves the users to understand better web usage data with the growth of Web-based applications, specifically e-commerce. A survey of association rule mining in text applications carried out by Manimaran et al. in [10]. This survey focuses on the use of Apriori algorithm and its application areas of ARM, TM and found suitable algorithm to analyze unknown information available in text data. After analyzing the work of various researchers, this survey concludes that among the different algorithms used to examine text data by using ARM technique, Apriori algorithm is suitable and mostly utilized in their chosen domains.

Another work carried out by Manimaran et al. in [11]. In this paper, they analyzed about the interestingness measures in ARM, this research takes only eight measures for analyzing their distinct feature. In this work, interesting measures has been calculated statistically within the resulting rules of apriori algorithm. They conclude that this research is to identify some interesting correlation measures as lift, chi-squared, hyper-lift, hyper-confidence and conviction. Qiaozhu Mei et al. carried out a research work by a review paper titled as "Discovering Evolutionary Theme Patterns from Text an Exploration of Temporal Text Mining" [17]. This paper is analyzed to discover the evolutionary theme pattern, they proposed a hidden Markov models to analyze the life cycle of each theme. This method not only used to find the strength variations of themes and also compares the relative strength of different themes.

Another work carried out by Turney and Peter in Mining the Web for Synonyms: PMI-IR Versus LSA on TOEFL in [24]. This paper is tells about the simple unsupervised learning algorithm for recognizing synonyms and it measures the semantic similarities. The algorithm is evaluated using the Test

of English as a Foreign Language. The algorithm is compared with Latent Semantic Analysis, which has also been evaluated using TOEFL. The comparison results that it suggesting several new hypotheses that are worth investigating. Andreas Hotho et al. carried out an article titled as A Brief Survey of Text Mining using data mining techniques [2]. In this article, they tell about the enlarged field of text mining. This survey paper briefly explains to find a specific problem with the currently available text mining algorithms, using their properties and their applications.

A research work titled as Survey of Text Mining: Clustering, Classification, and Retrieval is carried out by Michael W. Berry et al. in [15]. In this paper, they focused on the different level of complexity using document representation model and different document representation techniques. They have demonstrated that in the context of the classification, the better document representation produces better classification. Satoshi Morinaga et al. carried out in Mining Product Reputations on the Web in [23]. In this paper, they have explored a framework for mining product reputations on the web. It consists of an opinion extraction portion and a text mining portion as an application specific, question answering systems, and latter conducts four fundamental tasks: characteristics word extraction, co-occurring word extraction, typical sentence extraction and correspondence analysis. The key to combining these two parts is opinion labeling, which makes it possible to conduct supervised learning in the text-mining portion. Another work is carried out by Raymond Kosala and Hendrik Blockeel in Web Mining Research: A Survey in [19]. In this research paper, they survey the confusion regarding the usage in the area of Web mining. They also advise three Web mining categories and they also explore the connection between Web mining categories and the related agent paradigm. Vishal Gupta et al. carried out in A Survey of Text Mining Techniques and Applications in [24]. In this article, they suggest that in research most important area is Text Mining. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.

IV. THE ROLE OF CLASSIFICATION ALGORITHMS IN TEXT MINING

The classification algorithms are effectively utilized by various peoples for the real world applications. One of the major areas is exactly used in text mining applications including in the World Wide Web for different kind of analysis of text. In this section, it is discussed about the use of various classification algorithms, particularly ID3, C4, C4.5 and C5.0. This includes the real time marketing survey, text classification, e-mail classifications and auxiliary information using classification algorithms.

A research work titled as a personalized recommender system based on web usage mining and decision tree induction is carried out by Y.H. Cho et al. in [30]. They explored a methodology for e-commerce for developing a recommender system using decision tree induction. For further researches, it is interesting to compare the suggested methods with existing methods and it is very interesting to perform real time marketing survey to customers using this suggested method

and assess the performance. A research work by Michael J. Pazzani and Daniel Billsus in Content-Based Recommendation Systems in [14]. This article tells that the users can't get the content-based recommendation system, if the data is not containing enough information. For example, we can't differentiate in finding lawyer jokes and funny lawyer jokes. In such situations, collaborative recommenders should be used.

Monika D Khatri et al. carried out in Implementation of Text Mining with auxiliary Information using classification in [16]. In this paper, they proposed a software to find the required information like links and additional information which are commonly used in braces like (), { }, [] etc. If they want to find the manuscript enclosed in any type of braces, they have used WordNet 2.1 for obtaining the meaning of input words. Hence, the software is humble to use and deliver effectual output.

The table 1 shows that the summary of various methods proposed by different researchers in classifying the text in data mining. Most of the proposed methods are integration of different techniques and tells the conclusions and future work of the proposed algorithms, techniques and methods. Also from table 1, it is observed that the algorithms utilized for classifying the text data are performed well based on its input and approaches. Particularly, among the classification algorithms, the performance of C4.5 stamps its superiority in terms of accuracy and efficiency.

Table 1: Result Comparison

Paper Ref. No.	Author Name	Methods used	Results		
			Accuracy		
[29]	Wenmin Li et al.	C4.5, CBA and CMAR	CBA	C4.5	CMAR
			84.69 %	83.34 %	85.22 %
[7]	Govindarajan . M	PC5.0 and OC5.0	PC5.0		OC5.0
			0.5%		7.2%
[4]	Brijain R Pat et al.	ID3, C4.5, C5.0	classifying speed, strong learning ability and simple construction		
[12]	Manish Mehta et al.	C4, CART, SLIQ	They conclude that SLIQ produces accurate decision tree than the C4 and magnitude faster than CART		
[21]	S. B. Kotsiantis	SVMs	To reduce a multi-class problem to a set of multiple binary classification problems.		
[18]	Rafael S. Parpinelli et al.	Ant-Miner and CN2	It is comparatively better predictive accuracy in Ant-Miner than CN2		
[3]	Bing Liu et al.	CARs	To solve the number of problems in the classification systems		
[5]	Chuntao Jiang et al.	W-gSpan	It constructs from the representation of graph and uses the concept as input for classification		

[24]	Turney et al.	PMI-IR Versus LSA on TOEFL	The comparison results that it suggesting several new hypotheses that are worth investigating.
[17]	Qiaozhu Mei et al.	Hidden Markov models	It compares the relative strength of different themes
[11]	Manimaran et al.	ARM	To identify some interesting correlation measures as lift, chi-squared, hyper-lift, hyper-confidence and conviction
[10]	Manimaran et al.	Apriori algorithm	To examine text data by using ARM technique
[22]	Sakurai et al.	IDTF and SVM	IDTF gave high precision ratios in the classification than SVM
[25]	Ur-Rahman et al.	C4.5, KNN, Naïve Bayes and SVM	In KNN, Naïve Bayes and SVM, MKTPKS gave better accuracies than simple term classification model

A research work titled as Text Classification using Graph Mining-based Feature Extraction is carried out by Chuntao Jiang et al. in [5]. This paper explains the text classification using a graph-based representation for both the content and structure of manuscript. A proposed mechanism for weighted subgraph mining is W-gSpan. It results in constructs from the representation of graph and uses the concept as input for classification. Another research work carried out by Fabrizio Sebastiani in [6]. In this paper, the automated classification of texts into predefined categories, in research community the dominant approach is based on a general inductive process automatically builds a classifier by learning from a set of predefined manuscripts. This survey discusses the main approaches to text classification that fall within the machine-learning paradigm.

Adrian S Barb and Chi-Ren Shyu in Visual-Semantic Modeling in Content-Based Geospatial Information Retrieval Using Associative Mining Techniques in [1]. They have established an innovative tactic to associate low-slung image features with high-slung semantics using novel association-rule mining techniques and decision tree construction. Govindarajan. M carried out in Text mining technique for data mining application in [7]. He proposed one text mining classifier method C5.0 for the hypothyroid dataset. The accuracy of proposed C5.0 (PC5.0) is 0.5% and for original C5.0 (OC5.0) is 7.2% and it is 6.7% more accurate than the original C5.0. Then he concludes that the proposed C5.0 classifier method is more accurate than original C5.0.

A research work carried by Sakurai et al. in an e-mail analysis method based on text mining techniques in [22]. This article proposes a new method called IDTF to analyze e-mails by text mining techniques. They applied 3 tasks: product, contents and address analysis task and it showed IDTF gave high precision ratios in the classification than SVM. Another research work carried out by Ur-Rahman et al. in [25]. They proposed a

method called MKTPKS based classification model and compared with simple term classification model. They used 4 classification models namely C4.5, KNN, Naïve Bayes and SVM. In KNN, Naïve Bayes and SVM, MKTPKS gave better accuracies than simple term classification model.

V. CONCLUSION

Data classification is now a common task applied in many application areas such as grouping similar functional genomes, text that demonstrate the same pattern, partitioning web pages showing the same structure, and so on. The ID3 and C4.5 is the most well-known algorithm commonly used for classifying the data. This research work addresses various methods, techniques and performance of Classification Algorithms in Text mining. It is not possible to predict and suggest the best algorithms for any kind of applications in data mining because, the results are differ from one application to another application. But, from the various researchers' perspective, the classification algorithm C4.5 outperformed than the other algorithms in classifying the text data.

REFERENCES

- [1] Adrian S Barb and Chi-Ren Shyu, "Visual-Semantic Modeling in Content-Based Geospatial Information Retrieval Using Associative Mining Techniques", IEEE Geoscience and remote sensing letters, Vol. 7, No. 1, 2010, pp. 38-42.
- [2] Andreas Nürnberger, Hotho A., and Gerhard Paaß, "A Brief Survey of Text Mining", In Ldv Forum, Vol. 20, No. 1, 2005, pp. 19-62.
- [3] Bing Liu, Wynne Hsu and Yiming Ma, "Integrating Classification and Association Rule Mining", Proceedings of KDD-98, 1998, <http://www.aaai.org/Papers/KDD/1998/KDD98-012.pdf>.
- [4] Brijain R Patel and Kushik K Rana, "A Survey on Decision Tree Algorithm for Classification", International Journal of Engineering Development and Research, Vol. 2, Issue 1, 2014, pp. 1-5.
- [5] Chuntao Jiang, Frans Coenen, Robert Sanderson, and Michele Zito, "Text Classification using Graph Mining-based Feature Extraction", Knowledge-Based Systems, Vol. 23, No. 4, 2010, pp. 302-308.
- [6] Fabrizio Sebastiani, "Machine Learning in Automated Text Categorization", ACM computing surveys (CSUR), Vol. 34, No. 1, 2002, pp. 1-47.
- [7] Govindarajan. M, "Text mining technique for data mining application", In Proceedings of world academy of science, engineering and technology, Vol. 26, No. 104, 2007, pp. 544-549.
- [8] Ian H. Witten and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", SIGMOD Record, Vol. 31, No. 1, 2002, pp. 76-77.
- [9] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, Issue 2, 2000, pp. 1-12.
- [10] Manimaran. J and T. Velmurugan, "A survey of association rule mining in text applications", In Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, 2013, pp.698-702.
- [11] Manimaran. J and T. Velmurugan, "Analysing the quality of Association Rules by Computing an Interestingness Measures", Indian Journal of Science and Technology, Vol. 8(15), IPL094, 2015, pp.1-12.
- [12] Manish Mehta, Rakesh Agrawal and Jorma Rissanen, "SLIQ: A Fast Scalable Classifier for Data Mining", IBM Almaden Research Center, In Advances in Database Technology-EDBT'96, Springer Berlin Heidelberg, 1996, pp. 18-32.
- [13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, "Spatial Data Mining: A Database Approach", Institute for Computer Science, In Advances in spatial databases, Springer Berlin Heidelberg, 1997, pp. 47-66.
- [14] Michael J. Pazzani and Daniel Billsus, "Content-Based Recommendation Systems", The Adaptive Web, LNCS 4321, Springer-Verlag Berlin Heidelberg, 2007, pp. 325 – 341.
- [15] Michael W. Berry and Malu Castellanos, "Survey of Text Mining: Clustering, Classification, and Retrieval", Springer, 2007, pp. 1-247.
- [16] Monika D Khatri, S.S Dhande, "Implementation of Text Mining with auxiliary Information using classification", International Journal for Technological Research in Engineering, Vol. 2, Issue 10, 2015, pp. 2387-2393.
- [17] Qiaozhu Mei and Cheng Xiang Zhai, "Discovering Evolutionary Theme Patterns from Text an Exploration of Temporal Text Mining", In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 198-207.
- [18] Rafael S. Parpinelli, Heitor S. Lopes and Alex A. Freitas, "Data mining with an ant colony optimization algorithm", Evolutionary Computation, IEEE Transactions, Vol. 6, No. 4, 2002, pp. 321-332.
- [19] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, 2000, Vol. 2, Issue 1, pp. 1-15.
- [20] Robert Cooley, Bamshad Mobasher and Jaideep Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Knowledge and Information Systems, Springer Verlag, 1999, pp. 5-32.
- [21] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Department of Computer Science and Technology, 2007, pp. 249-268.
- [22] Sakurai, Shigeaki, and Akihiro Suyama, "An e-mail analysis method based on text mining techniques", Applied Soft Computing, Vol. 6, No. 1, 2005, pp. 62-71
- [23] Satoshi Morinaga, Kenji Yamanishi, Kenji Tateishi and Toshikazu Fukushima, "Mining Product Reputations on the Web", SIGKDD'02, 2002, pp. 341-349.
- [24] Turney and Peter, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL", Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001), 2001, pp. 491-502.
- [25] Ur-Rahman, Nadeem, and Jennifer A. Harding, "Textual data mining for industrial knowledge management and text classification: A business oriented approach", Expert Systems with Applications, Vol. 39, No. 5, 2012, pp. 4729-4739.
- [26] Vassilios S. Verykios, Elisa Bertino, Igor Nai Fovino, "State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record, Vol. 33, No. 1, 2004, pp. 50-57.
- [27] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence, Vol. 1, No. 1, 2009, pp. 60-76.
- [28] Wenke Lee, Salvatore J. Stolfo and Kui W. Mok, "A Data Mining Framework for Building Intrusion Detection Models", Proceedings of the 1999 IEEE Symposium on, In Security and Privacy, 1999, pp. 120-132.
- [29] Wenmin Li, Jiawei Han and Jian Pei, "CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules", ICDM 2001, Proceedings IEEE International Conference on, In Data Mining, 2001, pp. 369-376.
- [30] Y.H. Cho, Jae Kyeong Kim and Soung Hie Kim, "A personalized recommender system based on web usage mining and decision tree induction", Expert Systems with Applications, 2002, pp. 329-342.