2009 International Conference on Machine Learning and Computing
IPCSIT vol.3 (2011) © (2011) IACSIT Press, Singapore

An Overview of E-Documents Classification

Aurangzeb Khan [+], Baharum B. Bahurdin, Khairullah Khan

Department of Computer & Information Science Universiti Teknologi, PETRONAS

Abstract. With the increasing availability of electronic documents and the rapid growth of the World Wide Web, the task of automatic categorization of documents becomes the key method for organizing the information, knowledge and trend detection. With the growing availability of online resources, and popularity of fast and rich resources on web, classification of e-documents, news, personal blogs, and extraction of knowledge and trend from the documents has become an interesting area for research, as the World Wide Web is the fastest media for news and events collection from world. So the growing phenomenon of the textual data needs text mining, machine learning and natural language processing techniques and methodologies to organize and extract pattern and knowledge from the documents. This overview focused on the existing literature and explored the main techniques and methods for automatic documents classification i.e. documents representation, classifier construction and knowledge extraction and also discussed the issues along with the approaches and opportunities.

Keywords: Text mining, Web mining, Documents classification, Information retrieval.

1. Introduction

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

Natural Language Processing, Data Mining, and Machine learning techniques work together to automatically discover pattern from the documents. Text mining deals the categories of operations, retrieval, classification (supervised, unsupervised and semi supervised) summarization, trend and association analysis. The main goal of text mining is to enable users to extract information from textual resources. How the documented can be proper annotated, presented and classified, so the documents categorization consist several challenges, proper annotation to the documents, appropriate document representation, an appropriate classifier function to obtain good generalization and avoid over-fitting, also an appropriate dimensionality reduction method need to handle algorithmic issues [1].

Today web is the main resource for the text documents. The amount of textual data available to us is consistently increasing, according to [2] [3] approximately 80% of the information of an organization is stored in unstructured textual form in the form of reports, email, views and news. Information intensive business processes demand that we transcend from simple document retrieval to “knowledge” discovery. The need of automatically extraction of useful knowledge from the huge amount of textual data in order to assist the human analysis is fully apparent [4]. Market Trends based on the content of the online news articles, sentiments, and events is an emerging topic for research in data mining and text mining community. [5] [6]. This paper covers the overview of syntactic and semantic matters, domain ontology, tokenization concern and focus on the different machine learning techniques for text representation, categorization and knowledge detection from the text documents in the existing literature.

[+] Corresponding author
E-mail address: aurangzeb_khan@yahoo.com

In this overview we have used system literature review process. We have followed standard steps for searching, screening, data-extraction, and reporting. We tried to search for relevant paper, presentations, and research reports of the text mining using IEEE Explore, Springer Linker, Science Direct, ACM Portal and Googol Search Engine. The search is conducted after 1980, Published and/or unpublished research, focus on document mining, Machine Learning and Natural Language Processing (NLP). The non English writing and study before 1980 were excluded. Due to limited no of pages required, the detail explanations are not included.

The rest of the paper is organized as follows. In Section 2 we have given an overview of documents representation approaches, Section 3 presents document classification models, In section 4 opportunities and issues are discuss and also concludes the paper.

2. Document Representation

The document representation is the preprocessing process that is used to reduce the complexity of the documents and make them easier to handle, which needs to be transformed from the full text version to a document vector. Text representation is the important aspect in documents categorization that denotes the mapping of a document into a compact form of its content. A major characteristic of the text classification problem is the extremely high dimensionality of text data, so the number of potential features often exceeds the number of training documents. The documents have to be transformed from the full text version to a document vector which describes the contents of the document. Text classification is an important component in many informational management tasks, however with the explosive growth of the web data, algorithms that can improve the classification efficiency while maintaining accuracy, are highly desired [7].Dimensionality reduction (DR) is a very important step in text categorization, because irrelevant and redundant features often degrade the performance of classification algorithms both in speed and classification accuracy, the current literature shows that lot of work are in progress in the pre-processing and DR, and many models and techniques have been proposed. DR techniques can classify into Feature Extraction (FE) approaches and feature Selection (FS), as discussed bellow.

2.1. Feature Extraction

The process of feature extraction is to make clear the border of each language structure and to eliminate as much as possible the language dependent factors, tokenization, stop words removal, and stemming [9]. Feature Extraction is fist step of pre processing which is used to presents the text documents into clear word format. Removing stops words and stemming words is the pre-processing tasks.[10]. The documents in text classification are represented by a great amount of feature and most of then could be irrelevant or noisy [8]. Dimension reduction is the exclusion of a large number of keywords, base preferably on a statistical criterision, to create a low dimension vector [11]. Dimension Reduction techniques have attached much attention recently science effective dimension reduction make the learning task such as classification more efficient and save more storage space [12]. Commonly the steeps taken please for the feature extractions (Fig-1) are: Tokenization: A document is treated as a string and then partitioned into a list of tokens. Removing stop words: Stop words such as “the”, “a”, “and”… etc are frequently occurring, so the insignificant words need to be removed. Stemming word: Applying the stemming algorithm that converts different word form into similar canonical form. This step is the process of conflating tokens to their root form eg. Connection to connect, computing to compute etc.

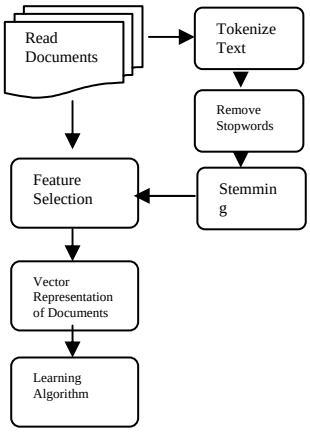

Figure 1: Document Classification

2.2. Feature Selection

After feature extraction the important step in pre-processing of text classification, is feature selection to construct vector space or bag of words, which improve the scalability, efficiency and accuracy of a text classifier. In general, a good feature selection method should consider domain and algorithm characteristics [13]. The main idea of FS is to select subset of feature from the original documents. FS is performed by keeping the words with highest score according to predetermined measure of the importance of the word [8]. Hence feature selection is commonly used in text classification to reduce the dimensionality of feature space and improve the efficiency and accuracy of classifiers.

There are mainly two types of feature selection methods in machine learning; wrappers and filters. Wrapper are much more time consuming especially when the number of features is high. As opposed to wrappers, filters perform feature selection independently of the learning algorithm that will use the selected features. In order to evaluate a feature, filters use an evaluation metric that measures the ability of the feature to differentiate each class [14]. We need to find the best matching category for the text document. The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weight each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories. Some of the recent literature shows that works are in progress for the efficient selection of the feature selection to optimize the classification process. A new feature selection algorithm is presented in [15], that is based on ant colony optimization to improve the text categorization. The authors in [16] developed a new feature scaling method, called class–dependent–feature–weighting (CDFW) using naive Bayes (NB) classifier. Many feature evaluation metrics have been explored, notable among which are information gain (IG), term frequency, Chi-square, expected cross entropy, Odds Ratio, the weight of evidence of text, mutual information, Gini index,(Table-1). A good feature selection metric should consider problem domain and algorithm characteristics. The authors in [19] focus on document representation and demonstrate that the choice of document representation has a profound impact on the quality of the classifier. In [18] the authors present significantly more efficient indexing and classification of large document repositories, e.g. to support information retrieval over all enterprise file servers with frequent file updates.

Table -1: Feature Selection Metrics

Gain Ration	$GR(t_k, c_i) = \dfrac{\sum_{c \in \{c_i, \bar{c}_i\}} \sum_{t \in \{t_k, \bar{t}_k\}} p(t,c) \log \dfrac{P(t,c)}{P(t)P(c)}}{-\sum_{c \in \{c_i, \bar{c}_i\}} P(c) \log P(c)}$
Informational Gain	$IG(w) = -\sum_{j=1}^{K} P(c_j) \log P(c_j) + P(w) \sum_{j=1}^{K} P(c_j \mid w) \log P(c_j \mid w) + P(\bar{w}) \sum_{j=1}^{K} P(c_j \mid \bar{w}) \log P(c_j \mid \bar{w})$ $= H(samples) - H(samples \mid w)$
Chi Square	$\chi^2(f_{i,}c_j) = \dfrac{\lvert D \rvert \times (\#(c_j, f_{i,})\#(\bar{c}_j, \bar{f}_{i,}) - \#(c_j, \bar{f}_{i,}).\#(\bar{c}_j, f_{i,}))^2}{(\#(\#(c_j, f_{i,}) + \#(c_j, \bar{f}_{i,})) \times (\#(\bar{c}_j, f_{i,}) + \#(\bar{c}_j, \bar{f}_{i,})) \times ((c_j, f_{i,}) + \#(\bar{c}_j, f_{i,})) \times (\#(c_j, \bar{f}_{i,}) + \#(\bar{c}_j, \bar{f}_{i,}))}$
Conditional mutual Information	$CMI(C \mid S) = H(C) - H(C \mid S_1, S_2, ..., S_n)$
Document Frequency	$DF(t_k) = P(t_k)$
Term Frequency	$tf(f_i, d_j) = \dfrac{freq_{ij}}{\max_k freq_{\kappa j}}$
Inverse Document Frequency	$\lvert idf \rvert = \log \dfrac{\lvert D \rvert}{\lvert \#(f1) \rvert}$
Term	$s(t) = P(t \in y \mid t \in x)$
Weighted Ration	$WOddsRatio\ n(w) = P(w) \times OddsRatio(w)$
Odd Ration	$OddsRatio(f_i, c_j) = \log \dfrac{P(f_i \mid c_j)(1 - P(f_i \mid \neg c_j))}{(1 - P(f_i \mid c_j))(P(f_i \mid \neg c_j)}$

2.3. Semantic and Ontology Based Documents Representation

Ontology is a data model that represents a set of concepts within a domain and the relationships between those concepts. It is used to reason about the objects within that domain. Ontology is the explicit and abstract model representation of already defined finite sets of terms and concept, involved in knowledge management, knowledge engineering, and intelligent information integration [19].Web Ontology Language (OWL) is the ontology support language derived from America DAPRA Agent Markup Language (DAML). Ontology has been proposed for handling semantically heterogeneity when extracting informational from various text sources such as internet [20].

Machine learning algorithms automatically builds a classifier by learning the characteristics of the categories from a set of classified documents, and then uses the classifier to classify documents into predefined categories. However, these machine learning methods have some drawbacks: (1) In order to train classifier, human must collect large number of training text term, the process is very laborious. If the predefined categories changed, these methods must collect a new set of training text terms. (2) Most of these traditional methods haven't considered the semantic relations between words. So, it is difficult to improve the

accuracy of these classification methods, (3) The issue of translatability, between one natural language into another natural language, identifies the types of issues that machine understanding systems are facing. These type of issues are discussed in the literature, some of these issues may be addressed if we have machine readable ontology [21], and that's why this is a potential area for research. During the text mining process, ontology can be used to provide expert, background knowledge about a domain. In [20] the author concentrates on the automatic classification of incoming news using hierarchical news ontology, based on this classification on one hand, and on the users' profiles on the other hand. A novel ontology-based automatic classification and ranking method is represented in [22] where Web document is characterized by a set of weighted terms, categories is represented by ontology. In [23] the author presented an approach towards mining ontology from natural language. In [25] the author presented a novel text categorization method based on ontological knowledge that does not require a training set. An automatic document classifier system based on Ontology and the Naïve Bayes Classifier is proposed in [26]. Ontology have shown their usefulness in application areas such as knowledge management, bioinformatics, e-learning, intelligent information integration, information brokering and natural-language processing and the positional and challenging area for text categorization.

Semantic analysis is the process of linguistically parsing sentences and paragraphs into key concepts, verbs and Proper Nouns. Using statistics-backed technology, these words are then compared to your taxonomy (categories) and grouped according to relevance. [27]. According to [28] the statistical techniques are not sufficient for the text mining. Better classification will be performed when consider the semantic under consideration, so the semantically representation of text and web document is the key challenge for the documents classification, knowledge and trend detection. The authors in [29] present the ambiguity issues in natural language text and present anew technique for resolving ambiguity problem in extracting concept/entity from the text which can improve the document classification. Multilingual text representation and classification is on of the main and challenging issue in text classification. In [24] the author presented the idea of workflow composition, and addressed the important issues of semantic description such as services for particular text mining task. Some of the natural language issues that should be consider during the text mining process shown in overview [30] is listed bellow in Table-2. Semantically representation of documents is the most challenging area for research in text mining. This will improve the classification and the information retrieval process.

Table -2: Semantic base text classification issues

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

3. Documents Classification

The documents can be classified by three ways unsupervised, supervised and semi supervised classification Many techniques and algorithms are proposed recently for the clustering and classification of electronic documents, our focus in these selection will be on the supervised classification techniques, new development and some future research direction from the existing literature. The automatic classification of documents into predefined categories has observed as an active attention as the Internet usage rate has quickly enlarged. From last few years , the task of automatic text categorization have been extensive study and rapid progress seems in this area, including the machine learning approaches such as Bayesian, Decision Tree and K-nearest neighbor(KNN) classifier. Support Vector Machines, Neural Networks, Latent Semantic Analysis, Rocchio's Algorithm, Fuzzy Correlation, Genetic Algorithms etc. Normally supervised learning techniques are used for automatic text categorization, where pre-defined category labels are assigned to documents based on the likelihood suggested by a training set of labeled documents.

Rocchio's Algorithm build prototype vector for each class i.e. the average vector over all training document vectors that belong to class C_i calculate similarity between test document and each of prototype vectors.

$$C_i = \alpha * centroid_{C_i} - \beta * centroid_{\overline{C_i}} \quad (1)$$

Assign test document to the class with maximum similarity this algorithm is easy to implement, fast learner and have relevance feedback mechanism but low classification accuracy, linear combination is too simple for classification and constant α and β are empirical. This is a widely used relevance feedback algorithm that operates in the vector space model [45]. The researchers have used a variation of Rocchio's algorithm in a machine learning context, i.e., for learning a user profile from unstructured text [46]. The goal in these applications is to automatically induce a text classifier that can distinguish between classes of documents

The k-nearest neighbor algorithm kNN [40] is used to test the degree of similarity between documents and k training data and to store a certain amount of classification data, thereby determining the category of test documents. In binary classification problems, it is helpful to choose k to be an odd number as this avoids tied votes and calculate similarity between test document and each neighbor assign test document to the class which contains most of the neighbors

.

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

This method is effective, non parametric and easy to implement, as compare to Racchio algorithm more local characteristic of documents are considered, however the classification time is long and difficult to find optimal value of k. The authors in [41] proposed a new algorithm incorporating the relationship of concept-based thesauri into document categorization using a kNN classifier. While [39] presents the use of phrases as basic features in the email classification problem. they performed extensive empirical evaluation using our large email collections using two k-NN classifiers with TF- IDF weighting and resemblance respectively.

Decision tree is also widely applied to document classification. The benefit is that it enables conversion into interpretable IF–THEN, and features rapid classification. [32] presents a hybrid method of rule-based processing and back-propagation neural networks for spam filtering. However, it can be shown experimentally that text classification tasks frequently involve a large number of relevant features [47]. Therefore, a decision tree's tendency to base classifications on as few tests as possible can lead to poor performance on text classification. However, when there are a small number of structured attributes, the performance, simplicity and understandability of decision trees for content-based models are all advantages. [48] describe an application of decision trees for personalizing advertisements on web pages. Decision tree is easy to understand and generate rules which reduce problem complexity but during implantation the training time is relatively expensive and does not handle continues variables also may suffer from overfitting.

Based on Bayes principle, Naïve Bayes is used to calculate the characteristics of a new document using keywords and joint probability of document categories and estimate the probability of each class for a document

$$P(c_i \mid D) = \frac{P(c_i)P(D \mid c_i)}{P(D)} \quad (3) \text{ , } \quad P(D \mid c_i) = \prod_{j=1}^{n} P(d_j \mid c_i) \quad (4)$$

Where P(Ci)= $P(C = c_i) = \frac{Ni}{N}$ and P(dj|ci) = $P(d_j \mid c_i) = \frac{1 + N_{ji}}{M + \sum_{k=1}^{M} N_{ki}}$. It has been one of the popular machine learning methods for many years. Its simplicity makes the framework attractive in various tasks and reasonable performances are obtained in the tasks although this learning is based on an unrealistic independence assumption. For this reason, there also have been many interesting works of investigating naive Bayes. Recently the [51][52] shows very good results by selecting Naïve bayes with SVM and SOM for text classification. The authors in [53] propose a Poisson naive Bayes text classification model with weight-enhancing method. Researcher shows great interest in naïve bayes classifier for spam filtering [54]. Naive Bayes work well on numeric and textual data, easy to implement and computation comparing with other algorithms however conditional independence assumption is violated by real-world data and perform very poorly when features are highly correlated and does not consider frequency of word occurrences.

In recent years, neural network has been applied in document classification systems to improve efficiency. Text categorization models using back-propagation neural network (BPNN) and modified back-propagation neural network (MBPNN) are proposed in [35]. Neural network based document classification methodology presented in [42] which is helpful for companies to manage patent documents more effectively.

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

Fuzzy correlation can deal with fuzzy information or incomplete data, and also convert the property value into fuzzy sets for multiple document classification [43]. The investigation in [36] explores the challenges of multi-class text categorization using one-against-one fuzzy support vector machine with Reuter's news as the example data, and shows better results using fuzzy correlation technique.

Genetic algorithm [49] aim to find optimum characteristic parameters using the mechanisms of genetic evolution and survival of the fittest in natural selection. Genetic algorithms make it possible to remove misleading judgments in the algorithms and improve the accuracy of document classification. This is an adaptive probability global optimization algorithm, which simulated in a natural environment of biological and genetic evolution, and they are widely used for their simple and strong. Now several researchers used this method of text classification. The authors in [50] introduced the genetic algorithm to text categorization and used to build and optimize the user template, and also introduced simulated annealing to improve the shortcomings of genetic algorithm. In the experimental analysis, they show that the improved method is feasible and effective for text classification.

Support Vector Machine (SVM) is supervised learning method for classification to find out the linear separating hyperplane which maximize the margin, i.e., the optimal separating hyperplane (OSH) and maximizes the margin between the two data sets. The authors in [33] implemented and measured the performance of the leading supervised and unsupervised approaches for multilingual text categorization and selected SVM as representative of supervised techniques. In [34] the authors analyses and compares SVM ensembles with four different ensemble constructing techniques. An optimal SVM algorithm via multiple optimal strategies is developed in [31]. [37, 38, 51, 52] presented latest technique for documents classification.

Among all the classification techniques SVM and Naïve Bayes has been recognized as one of the most effective and widely used text classification methods [44] provides a comprehensive comparison of supervised machine learning methods for text classification .

4. Discussion and Conclusion

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

Several algorithms or combination of algorithms as hybrid approaches are proposed for the automatics classification of documents, among these SVM and NB classifier are shown most appropriate in the existing literature. However more research is required for the performance improvement and accuracy of the documents classification and new method to solutions are required for useful knowledge from the increasing volume of electronics documents. The following are the some opportunities of the unstructured data classification and knowledge management.

- To reduce the training and testing time and improve the classification accuracy, precision, recall, micro-average macro-average.
- Spam filleting and email categorization: User may have folders like, electronic bills, email from family and friends, and so on, and may want a classifier to classify each incoming email and automatically move it to the appropriate folder. It is easier to find messages in sorted folders than in a very large inbox.
- Automatic allocation of folders to the downloaded articles, documents from text editors and from grid network.
- Semantic and Ontology: The use of semantics and ontology for the documents classification and informational retrieval.
- Trend mining i.e. marketing, business, and financial trend (stock exchange trend) form e-documents (Online news, stories, views and events).
- Mining text streams: Some new techniques and methods are required for handling stream text.

- Sentiments analysis: Automatic classification of sentiment and detection knowledge from it. So the sentiments and views mining is the new active area of text mining.
- Mining semi-structured documents: Classification and clustering of semi structured documents have some challenges and new opportunities.
- Words and Senses: An implementation of sense-based text classification, a procedure is needed for recovering the senses from the words used in a specific context.
- Information extraction: Informational extraction of useful knowledge from e- documents and Web pages, such as products and search results. to get meaning full patterns.

Welcome to The 2009 International Conference on Computer Engineering and Applications (ICCEA 2009). The Conference is a primary international forum for scientists and technicians working on topics relating to computer and applications. It will provide an excellent environment for participants to meet fellow academic professionals, to enhance communication, exchange and cooperation of most recent research, education and application on relevant fields. It will bring you new like-minded researchers, fresh idea. It also provides a friendly platform for academic and application professionals from crossing fields to communication together.

5. References

[1] A. Dasgupta, "Feature selection methods for text classification." *In Proceedings of the 13th ACMSIGKDD international conference on Knowledge discovery and data mining*, pp. 230 -239, 2007.

[2] Raghavan, P., S. Amer-Yahia and L. Gravano eds., "Structure in Text: Extraction and Exploitation." *In. Proceeding of the 7th international Workshop on the Web and Databases(WebDB): ACM SIGMOD/PODS* 2004, ACM Press, 2004, Vol 67.

[3] Oracle corporation, WWW,oracle.com, last visited 2008.

[4] Merrill lynch, Nov.,2000. *e-Business Analytics*: Depth Report.

[5] Pegah Falinouss "Stock Trend Prediction using News Article's: a text mining approach" *Master thesis* -2007.

[6] Andreas Hotho "A Brief Survey of Text Mining" , 2005.

[7] Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., and Wang Z., " A Noval Feature Selection Algorithm for text catogorization." *Elsevier, science Direct Expert system with application* , 33(1), pp.1-5, 2006.

[8] Montanes,E., Ferandez, J., Diaz, I., Combarro, E.F and Ranilla, J., " Measures of Rule Quality for Feature Selection in Text Categorization", *international Symposium on Intelligent data analysis* , Germeny-2003, Springer, Vol 28-10, pp.589-598, 2003.

[9] Wang, Y., and Wang X.J., " A New Approach to feature selection in Text Classification", *Proceedings of 4th International Conference on Machine Learning and Cybernetics*, IEEE- 2005, Vol.6, PP. 3814-3819, 2005.

[10] Lee, L.W., and Chen, S.M., "New Methods for Text CategorizationBased on a New Feature Selection Method a and New Similarity Measure Between Documents", *IEA/AEI*, France 2006.

[11] Manomaisupat, P., and Abmad k., " Feature Selection for text Categorization Using Self Orgnizing Map", 2nd International Conference on Neural Network and Brain 2005, *IEEE press* Vol 3, pp.1875-1880, 2005.

[12] Yan, J., Liu, N., Zhang, B., Yan, S., Chen, Z., Cheng, Q., Fan, W., and Ma, W.,"Optimal Orthogonal centroid Feature selection for Text Categorization." *International conference on Reserch and IR, ACM SIGIR, Barizal*, 2005, pp.122-129, 2005.

[13] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li "An Optimal Svm-Based Text Classification Algorithm" *Fifth International Conference on Machine Learning and Cybernetics, Dalian,* 13-16 August 2006.

[14] Hiroshi Ogura, Hiromi Amano, Masato Kondo "Feature selection with a measure of deviations from Poisson in text categorization" *Expert Systems with Applications* 36 -2009, pp 6826–6832, 2009.

[15] Mehdi Hosseinzadeh Aghdam, Nasser Ghasem-Aghaee, Mohammad Ehsan Basiri "Text feature selection using ant colony optimization", *Expert Systems with Applications* 36- 2009, pp 6843–6853, 2009.

[16] E. Youn, M. K. Jeong , "Class dependent feature scaling method using naive Bayes classifier for text datamining" *Pattern Recognition Letters* -2009.

[17] G. Forman, E. Kirshenbaum, "Extremely Fast Text Feature Extraction for Classification and Indexing", *Napa Valley California, USA. CIKM'08,* October 26–30, 2008.

[18] Mostafa Keikha, Ahmad Khonsari, Farhad Oroumchian, " Rich document representation and classification: An analysis" , *Knowledge-Based Systems* 22 , 2009, pp. 67–71, 2009.

[19] D.Fensel, "Ontologies: Silver Bullet for Knowledge Management and e-Commerce", *Springer Verlag, Berlin,* 2000.

[20] Lena Tenenboim, Bracha Shapira, Peretz Shoval "Ontology-Based Classification Of News In An Electronic Newspaper" International Conference "Intelligent Information and Engineering Systems" INFOS 2008, *Varna, Bulgaria,* June-July 2008.

[21] Mu-Hee Song, Soo-Yeon Lim, Dong-Jin Kang, and Sang-Jo Lee, "Automatic Classification of Web pages based on the Concept of Domain Ontology", *Proc. Of the 12th Asia-Pacific Software Engineering Conference,* 2005.

[22] Jun Fang, Lei Guo, XiaoDong Wang and Ning Yang "Ontology-Based Automatic Classification and Ranking for Web Documents" *Fourth International Conference on Fuzzy Systems and Knowledge Discovery* (FSKD 2007.

[23] Alexander Maedche and Ste_en Staab "Mining Ontologies from Text" *LNAI 1937, Springer-Verlag* – 2000, pp. 189-202

[24] M. Sarnovský, M. Parali "Text Mining Workflows Construction with Support of Ontologies" *6th International Symposium on Applied Machine Intelligence and Informatics- SAMI* 2008.

[25] Maciej Janik and Krys Kochut "Training-less Ontology-based Text Categorization" -2007.

[26] Yi-Hsing Chang, Hsiu-Yi Huang, "An Automatic Document Classifier System Based On Naïve Bayes Classifier And Ontology" *Seventh International Conference on Machine Learning and Cybernetics, Kunming,* July 2008.

[27] http://www.nstein.com/en/tme_intro.php-, last visited- 2008.

[28] Yah, A.s., Hirschman, L., and Morgan, A.A. "Evaluation of text data mining for databasecuration: lessons learned from the KDD challenge cup." *Bioinformatics* 19(supp.1), 2003, pp.i331-i339.

[29] H.M.Al Fawareh, S.Jusoh, W.R.S.Osman, "Ambiguity in Text Mining", IEEE-2008.

[30] A.Stavrianou, P. Andritsos, N. Nicoloyannis "Overview and semantic issues of text mining", *SIGMOD* ,2007, Vol.36,N03.

[31] Zi-Qiang Wang, Xia Sun, De-Xian Zhang, Xin Li "An Optimal Svm-Based Text Classification Algorithm" *Fifth International Conference on Machine Learning and Cybernetics, Dalian,* August 2006.

[32] Chih-Hung Wu , "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks", *Expert Systems with Applications,* pp. 4321–4330, 2009.

[33] Chung-Hong Lee a, , Hsin-Chang Yang , "Construction of supervised and unsupervised learning systems for multilingual text categorization" , *Expert Systems with Applications,* pp. 2400–2410 , 2009.

[34] Shi-jin Wang, Avin Mathew, Yan Chen , Li-feng Xi , Lin Ma, Jay Lee, "Empirical analysis of support vector machine ensemble classifiers", *Expert Systems with Applications,* pp. 6466–6476, 2009.

[35] Bo Yu, Zong-ben Xu, Cheng-hua Li ,"Latent semantic analysis for text categorization using neural network", *Knowledge-Based Systems* 21- pp. 900–904, 2008.

[36] Tai-Yue Wang and Huei-Min Chiang "One-Against-One Fuzzy Support Vector Machine Classifier: An Approach to Text Categorization", *Expert Systems with Applications,* doi: 10.1016/j.eswa.2009.

[37] Wen Zhang a, Taketoshi Yoshida a, Xijin Tang "Text classification based on multi-word with support vector machine" , *Knowledge-Based Systems* 21 -pp. 879–886, 2008.

[38] Youngjoong Ko a, Jungyun Seo, "Text classification from unlabeled documents with bootstrapping and feature projection techniques", *Information Processing and Management* 45 -,pp. 70–83, 2009.

[39] Matthew Changa, Chung Keung Poon_, Using Phrases as Features in Email Classification", *The Journal of Systems and Software* ,doi: 10.1016/j.jss, 2009.

[40] Tam, V., Santoso, A., & Setiono, R. , "A comparative study of centroid-based, neighborhood-based and statistical approaches for effective document categorization", *16th International Conference on Pattern Recognition,* 2002- 4, 235–238.

[41] Bang, S. L., Yang, J. D., & Yang, H. J. , “Hierarchical document categorization with k-NN and concept-based thesauri. Information Processing and Management”, pp. 397–406, 2006.

[42] Trappey, A. J. C., Hsu, F.-C., Trappey, C. V., & Lin, C.-I., “Development of a patent document classification and search platform using a back-propagation network”. *Expert Systems with Applications*, pp. 755–765, 2006.

[43] Que, H. -E. “Applications of fuzzy correlation on multiple document classification.Unpublished master thesis” , *Information Engineering epartment, Tamkang University*, Taipei, Taiwan-2000.

[44] Y.Yang,and X.Liu, “An re-examination of text categorization”, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Berkeley, pp.42-49,August 1999.

[45] Ittner, D., Lewis, D., Ahn, D.: Text Categorization of Low Quality Images. In: Symposium on Document Analysis and Information Retrieval, Las Vegas, NV –pp.301-315, 1995.

[46] Pazzani M., Billsus, D., “Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning* 27(3) pp.313-331, 1997.

[47] Joachims, T.: Text Categorization With Support Vector Machines: *Learning with Many Relevant Features. In: European Conference on Machine Learning, Chemnitz*, Germany -1998- pp. 137-142 , 1998.

[48] Kim, J., Lee, B., Shaw, M., Chang, H., Nelson, W.: Application of Decision-Tree Induction Techniques to Personalized Advertisements on Internet Storefronts. *International Journal of Electronic Commerce -2001*, 5(3) pp. 45-62, 2001.

[49] Wang Xiaoping, Li-Ming Cao., “Genetic Algorithm Theory, Application and Software[M].*XI'AN:Xi'an Jiaotong* Uni. Pr, 2002.

[50] ZHU Zhen-fang, LIU Pei-yu, Lu Ran, “Research Of Text Classification Technology Based On Genetic Annealing Algorithm” *IEEE*, 2008, 978-0-7695-3311-7, 2008.

[51] Dino Isa, Lam Hong lee, V. P Kallimani, R. RajKumar, “ Text Documents Preprocessing with the Bahes Formula for Classification using the Support vector machine”, *IEEE, TKDE*, Vol-20, N0-9 pp-1264-1272, 2008.

[52] Dino Isa,, V. P Kallimani Lam Hong lee, “Using Self Organizing Map for Clustering of Text Documents”, *Elsever* , Expert System with Applications-2008.

[53] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaeng, “Some Effective Techniques for Naive Bayes Text Classification”, *IEEE, TKDE*, Vol. 18, No. 11, , Pp- 1457- 1466 ,November 2006.

[54] Thiago S.Guzella, Walimir M. Caminhas “ A Review of machine Learning Approches to Spam Filtering”, *Elsever* , Expert System with Applications-2009.