A SURVEY ON TEXT CATEGORIZATION

S.Niharika[#1], V.Sneha Latha[*2], D.R.Lavanya[#3]

[#] *Department of computer science and technology, KL University*
Green Fields, Vaddeswaram, Guntur, AP
[*] *Asst professor, Department of CSE, KL University*
Green Fields, Vaddeswaram, Guntur, AP

***Abstract*— Now a day's managing a vast amount of documents in digital forms is very important in text mining applications. Text categorization is a task of automatically sorting a set of documents into categories from a predefined set. A major characteristic or difficulty of text categorization is high dimensionality of feature space. The reduction of dimensionality by selecting new attributes which is subset of old attributes is known as feature selection. Feature-selection methods are discussed in this paper for reducing the dimensionality of the dataset by removing features that are considered irrelevant for the classification. In this paper we discuss several approaches of text categorization, feature selection methods and applications of text categorization.**

***Keywords*— Text mining, text classification, feature selection**

I. INTRODUCTION

The capacity of storing data becomes enormous as the technology of computer hardware develops. So amount of data is increasing exponentially, the information required by the users become varies .actually users deal with textual data more than the numerical data. It is very difficult to apply techniques of data mining to textual data instead of numerical data. Therefore it becomes necessary to develop techniques applied to textual data that are different from the numerical data. Instead of numerical data the mining of the textual data is called text mining. Text mining [1] is procedure of synthesizing the information by analyzing relations, the patterns and rules from the textual data. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. Text mining is different from what are familiar with in web search. In search, the user is typically looking for something that is already known and has been written by someone else. The problem is pushing aside all the material that currently is not relevant to your needs in order to find the relevant information. In text mining, the goal is to discover unknown information, something that no one yet knows and so could not have yet written down. The functions [2] of the text mining are text summarization, text categorization and text clustering. The content of this paper is restricted to text categorization.

Text categorization (or text classification) is the assignment of natural language documents to predefined categories according to their content [3]. The set of categories is often called a controlled vocabulary. Text classification is the act of dividing a set of input documents into two or more classes where each document can be said to belong to one or multiple classes. Huge growth of information flows and especially the explosive growth of Internet promoted growth of automated text classification. Development of computer hardware provided enough computing power to allow automated text classification to be used in practical applications. The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques [4]: a general inductive process automatically builds a classifier by learning, from a set of pre classified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains.

Text classification is commonly used to handle spam emails, classify large text collections into topical categories, used to manage knowledge and also to help Internet search engines. A major characteristic of text categorization is high dimensionality of the feature space .the native feature space consists of hundreds of thousands of terms for even a moderate sized text collection. Various feature selection methods are discussed in this paper to overcome the problem of the high dimensionality. This survey also focuses on the various approaches and also the applications of text categorization.

II. TEXT CATEGORIZATION

Categorization involves identifying the main themes of a document by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a "bag of words." It does not attempt to process the actual information as information extraction does.

Rather, categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a thesaurus for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms.

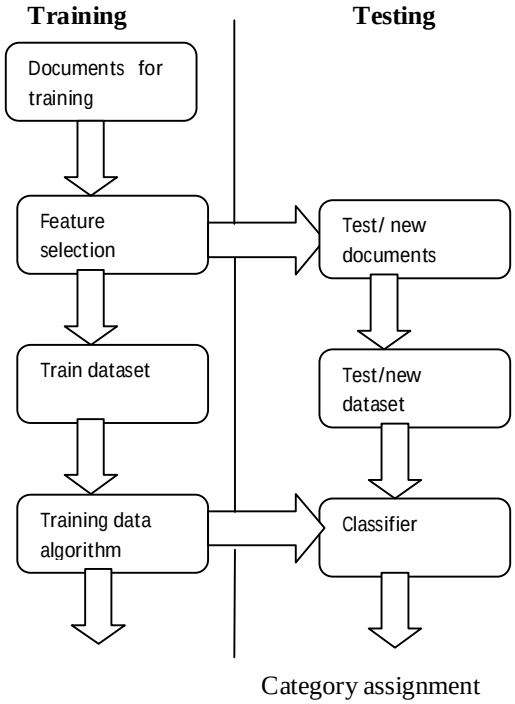

Fig. 1 process of text categorization

The goal of text categorization is to classify a set of documents into a fixed number of predefined categories. Each document may belong to more than one class. Using supervised learning algorithms [5], the objective is to learn classifiers from known examples (labeled documents) and perform the classification automatically on unknown examples (unlabeled documents). Figure.8 shows the overall flow diagram of the text categorization task. Consider a set of labeled documents from a source D = [d_1, d_2….d_n] belonging to a set of classes C= [c_1,c_2,…,c_p]. The text categorization task is to train the classifier using these documents, and assign categories to new documents. In the training phase, the *n* documents are arranged in *p* separate folders, where each folder corresponds to one class. In the next step, the training data set is prepared via a feature selection process. Next step, the training data set is prepared via a feature selection process.

Text data typically consists of strings of characters, which are transformed into a representation suitable for learning. It is observed from previous research that words work well as features for many text categorization tasks. In the feature space representation, the sequences of characters of text documents are represented as sequence of words. Feature selection involves tokenizing the text, indexing and feature space reduction. Text can be tokenized using term frequency (TF), inverse document frequency (IDF), term frequency inverse document frequency (TFIDF) or using binary representation. Using these representations the global feature space is determined from entire training document collection.

A. Single-label vs. multi-label text categorization

Different constraints may be enforced on the TC task, depending on the application. For instance we might need that, for a given integer k, exactly k (or $\leq$ k, or $\geq$ k) elements of C be assigned to each$d_j \in D$. The case in which exactly 1 category must be assigned to each $d_j \in D$. is often called the single-label text categorization. The assigning of number of categories from 0 to |c| to the same $d_j \in D$ is referred to as the multi-label text categorization. Multi-label text classification can be categorized into two different approaches .they are problem transformation methods and algorithm adaptation methods.

B. Category-pivoted vs. document-pivoted text categorization

There are two different ways of using a text classifier. Given$d_j \in D$, we might want to find all the $c_i \in C$ under which it should be filed (document-pivoted categorization– DPC); alternatively, given$c_i \in C$, we might want to find all the $d_j \in D$ that should be filed under it (category-pivoted categorization – CPC). DPC is used when documents become available at different moments in time, e.g. in filtering e-mail. CPC is instead suitable when a new category is added to an existing set after a number of documents have already been classified under C, and or when these documents need to be reconsidered for classification under new category (e.g. [Larkey 1999]).

III. CATEGORIZATION METHODS

A. Decision Trees

Decision tree methods rebuild the manual categorization of the training documents by constructing well-defined true/false-queries in the form of a tree structure where the nodes represent questions and the leaves represent the corresponding category of documents. After having created the tree, a new document can easily be categorized by putting it in the root node of the tree and let it run through the query structure until it reaches a certain leaf. The main advantage of decision trees is the fact that the output tree is easy to interpret even for persons who are not familiar with the details of the model [6]. The tree structure generated by the model provides the user with a consolidated view of the categorization logic and is therefore useful information. A risk of the application of tree methods is known as "over fitting": A tree over fits the training data if there exists

an alternative tree that categorizes the training data worse but would categorize the documents to be categorized later better. This circumstance is the result of the algorithm's intention to construct a tree that categorizes every training document correctly; however, this tree may not be necessarily well suited for other documents. This problem is typically moderated by using a validation data set for which the tree has to perform in a similar way as on the set of training data. Other techniques to prevent the algorithm from building huge trees (that anyway only map the training data correctly) are to set parameters like the maximum depth of the tree or the minimum number of observations in a leaf. If this is done, Decision Trees show very good performance even for categorization problems with a very large number of entries in the dictionary.

B. k-Nearest Neighbor

The categorization itself is usually performed by comparing the category frequencies of the k nearest documents (neighbors). The evaluation of the closeness of documents is done by measuring the angle between the two feature vectors or calculating the Euclidean distance between the vectors. In the latter case the feature vectors have to be normalized to length 1 to take into account that the size of the documents (and, thus, the length of the feature vectors) may differ. A doubtless advantage of the k-nearest neighbor method is its simplicity. It has reasonable similarity measures and does not need any resources for training. K nearest neighbor performs well even if the category-specific documents from more than one cluster because the category contains, e.g., more than one topic. This situation is badly suited for most categorization algorithms. A disadvantage is the above-average categorization time because no preliminary investment (in the sense of a learning phase) has been done. Furthermore, with different numbers of training documents per category the risk increases that too many documents from a comparatively large category appear under the k nearest neighbors and thus lead to an inadequate categorization.

C. Bayesian Approaches

There are two groups of Bayesian approaches in document categorization: Naïve [7] and non-naïve Bayesian approaches. The naïve part of the former is the assumption of word (i.e. feature) independence, meaning that the word order is irrelevant and consequently that the presence of one word does not affect the presence or absence of another one. A disadvantage of Bayesian approaches [8] in general is that they can only process binary feature vectors and, thus, have to abandon possibly relevant information.

D. Neural Networks

Neural networks consist of many individual processing units called as neurons connected by links which have weights that allow neurons to activate other neurons. Different neural network approaches have been applied to document categorization problems. While some of them use the simplest form of neural networks, known as perceptions, which consist only of an input and an output layer, others build more sophisticated neural networks with a hidden layer between the two others. In general, these feed-forward -nets consist of at least three layers (one input, one output, and at least one hidden layer) and use back propagation as learning mechanism. The advantage of neural networks is that they can handle noisy or contradictory data very well. The advantage of the high flexibility of neural networks entails the disadvantage of very high computing costs. Another disadvantage is that neural networks are extremely difficult to understand for an average user; this may negatively influence the acceptance of these methods.

E. Regression-based Methods

For this method the training data are represented as a pair of input/output matrices where the input matrix is identical to our feature matrix A and the output matrix B consists of flags indicating the category membership of the corresponding document in matrix A. Thus B has the same number of rows like A (namely m) and c columns where c represents the total number of categories defined. The goal of the method is to find a matrix F that transforms A into B' (by simply computing B'=A*F) so that B' matches B as well as possible. The matrix F is determined by applying multivariate regression techniques. An advantage of this method is that morphological preprocessing (e.g., word stemming) of the documents can be avoided without losing categorization quality. Thus, regression-based approaches become truly language-independent. Another advantage is that these methods can easily be used for both single category and multiple-category problems.

F. Vector-based Methods

We discuss two types of vector-based methods: The centroid algorithm and support vector machines. One of the simplest categorization methods is the centroid algorithm. During the learning stage only the average feature vector for each category is calculated and set as centroid-vector for the category. A new document is easily categorized by finding the centroid-vector closest to its feature vector. The method is also inappropriate if the number of categories is very large. Support vector machines (SVM) need in addition to positive training documents also a certain number of negative training documents which are untypical for the category considered. SVM is then looking for the decision surface that best separates the positive from the negative examples in the n-dimensional space. The document representatives closest to the decision surface are called support vectors. The result of the algorithm remains unchanged if documents that do not belong to the support

vectors are removed from the set of training data. An advantage of SVM [9] is its superior runtime-behavior during the categorization of new documents because only one dot product per new document has to be computed. A disadvantage is the fact that a document could be assigned to several categories because the similarity is typically calculated individually for each category.

IV. FEATURE SELECTION METHODS

Feature-selection methods play a very important role in the reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification [10]. These feature selection methods possess a number of advantages such as smaller dataset size, smaller computational requirements for the text categorization algorithms (especially those that do not scale well with the feature set size) and considerable shrinking of the search space. The goal is the reduction of the curse of dimensionality to yield improved classification accuracy. Another benefit of feature selection is its tendency to reduce overfitting, i.e. the phenomenon by which a classifier is tuned also to the contingent characteristics of the training data rather than the constitutive characteristics of the categories, and therefore, to increase generalization. Best Individual Features can be performed using some of the measures, for instance, document frequency, term frequency, mutual information, information gain, odds ratio, $\chi 2$ statistic and term strength [11], [12], [13], [14], [15]. What is common to all of these feature-scoring methods is that they conclude by ranking the features by their independently determined scores, and then select the top scoring features.

A. Document frequency:

Document frequency is number of documents in which a term occurs. DF thresholding is the simplest technique for the vocabulary reduction. We have to compute document frequency for each unique term in the training set and we have to discard all the terms whose frequency is less than the threshold value from the feature space. The removal of the rare terms reduces the dimensionality of the feature space.

B. Information gain

Information gain is frequently employed as a term goodness criterion in machine learning. The prediction of category is done by measuring by measuring number of bits of information and by knowing presence or absence of a term in the document. The information gain of term t is defined to be

$$G(t) = -\sum_{i=1}^{m} p_r(c_i)\log p_r(c_i) + p_r(t)\sum_{i=1}^{m} p_r(c_i/t)\log p_r(c_i/t) + p_r(t)\sum_{i=1}^{m} p_r(c_i/t)\log p_r(c_i/t)$$

Given a training corpus for each unique term we compute the information gain and remove from the feature space those terms whose information gain was less than some predetermined threshold. The computation includes the estimation of the conditional probabilities of a category given a term and the entropy computations in the definition.

C. Mutual information

Mutual information is a criterion commonly used in statistical language modeling of, word associations and related applications [16].This is able to provide a precise statistical calculation that could be applied to a very large corpus to produce a table of association of words. If one considers a two way contingency table of a term t and a category c. where A is number of times c and t co-occur. B is the number of times t occur without c. C is number of times c occur without t, N is the total number of documents, then mutual information criterion between t and c is defined to be

$$I(t, c) = \log \frac{P_r(t\hat{}c)}{P_r(t) \times P_r(c)}$$

I (t, c) has a natural value of zero if t and s are independent. A weakness of the mutual information is that score is strongly influenced by the marginal probabilities of the terms.

D. Term strength

Term strength is originally proposed and evaluated by Wilbur and sirotkin [17] for vocabulary reduction in text retrieval and later applied by yang and Wilbur to text categorization. This method estimates term importance based on how commonly a term is likely to appear in closely related documents. It uses a training set of documents to derive document pairs whose similarity is above threshold. Let x and y be arbitrary pair of distinct but related documents and t may be a term therefore the term strength may be defined as

$$S(t) = P_r(t \in y | t \in x)$$

E. X^2statistic

The X^2statistic measures the lack of independence between t and c and can be compared to the X^2distribution with one degree of freedom to judge extremeness. Here A is defined as number of times t and c co-occur. B is number of times the t occurs without c. C is number of times c occurs without t. D is number of times either c or t occurs and N is the total number of documents. It is defined as

$$X^2(t, c) = \frac{\boldsymbol{N \times (AD - BC)^2}}{\boldsymbol{(A + C) \times (B + D) \times (A + B) \times (C + D)}}$$

V. APPLICATIONS OF TEXT CATEGORIZATION

The applications of text categorization are manifold. Common traits among all of them are

- The need to handle and organize documents in which the textual component is either unique, or dominant, or simplest to interpret component.
- The need to handle and organize large quantities of such documents, i.e large enough that their manual organization into classes is either too expensive or not feasible within the time constraints imposed by the application.
- The fact that the set of categories is known in advance, and is variation over time is small.

A. Document Organization

A document organization is a collection of documents composed of labeled clusters that contain similar documents. Note that a collection of non-clustered documents is not a document organization. If the document organization contains clusters with nested clusters, it is called a hierarchical document organization. If its clusters do not have any nested clusters, it is called a flat document organization. It is necessary to build a document organization, manually or automatically, for the efficient management of documents. There are two types of document organizations, static document organization and dynamic document organization. If the clusters of the document organization are fixed permanently, it is called a static document organization. If it adapts by itself, to the current situation, we refer to the document organization as a dynamic document organization.

Indexing with a controlled vocabulary is an instance of the general problem of document base organization. For instance, at the offices of a newspaper incoming "classified" ads must be, prior to publication, categorized under categories such as Personals, Cars for Sale, Real Estate, etc. Newspapers dealing with a high volume of classified ads would benefit from an automatic system that chooses the most suitable category for a given ad. Other possible applications are the organization of patents into categories for making their search easier [18], the automatic filing of newspaper articles under the appropriate sections (e.g., Politics, Home News, Lifestyles, etc.), or the automatic grouping of conference papers into sessions.

B. Text Filtering

Text filtering is the activity of classifying a stream of incoming documents dispatched in an asynchronous way by an information producer to an information consumer [Belkin and Croft 1992]. A typical case is a newsfeed filter [19], where the producer is a news agency and the consumer is a newspaper. In this case, the filtering system should block the delivery of the documents the consumer is likely not interested in (e.g., all news not concerning sports, in the case of a sports newspaper). Filtering can be seen as a case of single-label TC, that is, the classification of incoming documents into two disjoint categories, the relevant and the irrelevant additionally, a filtering system may also further classify the documents deemed relevant to the consumer into thematic categories; in the example above, all articles about sports should be further classified according to which sport they deal with, so as to allow journalists specialized in individual sports to access only documents of prospective interest for them. Similarly, an e-mail filter might be trained to discard "junk" mail [20] and further classify non junk mail into topical categories of interest to the user. A filtering system may be installed at the producer end, in which case it must route the documents to the interested consumers only, or at the consumer end, in which case it must block the delivery of documents deemed uninteresting to the consumer.

C. Word Sense Disambiguation

Word ambiguity is not something that we encounter in everyday life, except perhaps in the context of jokes. Somehow, when an ambiguous word is spoken in a sentence, we are able to select the correct sense of that word without considering alternative senses. However, in any application where a computer has to process natural language, ambiguity is a problem. For example, if a language translation system encountered the word 'bat' in a sentence, should the translator regard the word as meaning: an implement used in sports to hit balls; or a furry, flying mammal?

Word sense disambiguation (WSD) is the activity of finding, given the occurrence in a text of an ambiguous (i.e., polysemous or homonymous) word, the sense of this particular word occurrence. For instance, bank may have (at least) two different senses in English, as in the Bank of England (a financial institution) or the bank of river Thames (a hydraulic engineering artifact). It is thus a WSD task to decide which of the above senses the occurrence of bank in Last week I borrowed some money from the bank has. WSD is very important for many applications, including natural language processing, and indexing documents by word senses rather than by words for IR purposes. WSD may be seen as a TC task (see [21]) once we view word occurrence contexts as documents and word senses as categories. Quite obviously, this is a single-label TC case, and one in which document-pivoted

D. Hierarchical categorization of Web pages

The Internet, mainly the World Wide Web and the Usenet, offers a lot of information to the interested user. The number of documents accessible via the net is growing rapidly. To manage

this chaotic state, engines like AltaVista1 or Yahoo!2 offer mechanisms to search for the documents that the user needs. Some of them, like AltaVista, let the user type in keywords describing the desired document. Others, like Yahoo!, put the documents into a hierarchically ordered category scheme so that the user can browse through these categories to satisfy his information needs. Categorization of web documents (e.g. HTML documents) denotes the task of finding relevant categories for a (new) document which is to be inserted into such a web catalogue. This is mostly done manually. But the large number of new documents which appear on the World Wide Web and need to be categorized raises the question of whether and how this task can be performed automatically.

Automatic categorization of web documents (e.g. HTML documents) denotes the task of automatically finding relevant categories for a (new) document which is to be inserted into a web catalogue like Yahoo!. There exist many approaches for performing this difficult task. Here, special kinds of web catalogues, those whose category scheme is hierarchically ordered, are regarded. A method for using the knowledge about the hierarchy to gain better categorization results is discussed. This method can be applied in a post-processing step and therefore be combined with other known (non-hierarchical) categorization approaches.

E. Spam filtering

The unwanted form of an email message is defined as spam. Filtering spam is a task of increased applicative interest that lies at cross roads between filtering and generic classification. In fact it has the dynamic character of other filtering applications such as email filtering, and it cuts across different topics, as genre classification. Several attempts some of them quite successful have been made at applying standard text classification techniques to spam filtering for applications involving either personal mail[22]or mailing lists[23].

One of the problems of spam filtering is unavailability of negative training messages. A software maker wishing to customize its spam filter for a particular client needs training examples, while positive ones (i.e. spam messages) are not hard to collect in large quantities, negative ones (i.e. legitimate messages) even to someone even to someone who is going to use these messages. Here we have to use the machine learning methods that do not use negative training examples.

F. Automatic survey coding

Survey coding is the task of assigning a symbolic code from a predefined set of such codes from a predefined set of such codes to the answer that a person has given in response to an open ended question in a questionnaire. This task is usually carried out to group respondents according to a predefined scheme based on their answers. Survey coding is a difficult task, since that the code that should be attributed to a respondent based on the answer she has given is a matter of subjective judgment, and thus requires expertise. The problem can be formulated as a single-label text categorization problem [24] where the answers play the role of the documents and the codes that are applicable to the answers returned to a given question play the role of categories.

VI. CONCLUSION

Text categorization play a very important role in information retrieval, machine learning , text mining and it have been successful in tackling wide variety of real world applications. Key to this success have been the ever-increasing involvement of the machine learning community in text categorization, which has lately resulted in the use of the very latest machine learning technology within text categorization applications. Many approaches for text categorization are discussed in this paper. Feature selection methods are able to successfully reduce the problem of dimensionality in text categorization applications. Process of text classification is well researched, but still many improvements can be made both to the feature preparation and to the classification engine itself to optimize the classification performance for a specific application. Research describing what adjustments should be made in specific situations is common, but a more generic framework is lacking. Effects of specific adjustments are also not well researched outside the original area of application. Due to these reasons, design of text classification systems is still more of an art than exact science.

REFERENCES

[1] Berry Michael W., Automatic Discovery of Similar Words, in “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, New York, LLC, 2004, pp.24-43.

[2] Vishal gupta and Gurpreet S. Lehal , “A survey of text mining techniques and applications”, journal of emerging technologies in web intelligence, 2009,pp.60-76.

[3] Sebastiani F., “Machine Learning in Automated Text Categorization”, ACM Computing Surveys, vol. 34 (1),2002, pp. 1-47.

[4] Zu G., Ohyama W., Wakabayashi T., Kimura F., "Accuracy improvement of automatic text classification based on feature transformation": Proc: the 2003 ACM Symposium on Document Engineering, November 20-22, 2003,pp. 118-120.

[5] Setu Madhavi Namburu, Haiying Tu, Jianhui Luo and Krishna R. Pattipati , “Experiments on Supervised Learning Algorithms for Text Categorization”, International Conference , IEEE computer society,2005, 1-8.

[6] D. E. Johnson, F. J. Oles, T. Zhang, T. Goetz,“A decision-tree-based symbolic rule induction system for text categorization”, IBM Systems Journal, September 2002.

[7] Kim S. B., Rim H. C., Yook D. S. and Lim H. S., “Effective Methods for Improving Naïve Bayes Text Classifiers”, LNAI 2417, 2002, pp.414-423.

[8] Klopotek M. and Woch M., “Very Large Bayesian Networks in Text Classification”, ICCS 2003, LNCS 2657, 2003, pp. 397-406.

[9] Joachims, T., Transductive inference for text classification using support vector machines. Proceedings of ICML-99, 16th International Conference on

Machine Learning, eds. I. Bratko & S. Dzeroski, Morgan Kaufmann Publishers, San Francisco, US: Bled, SL, ,1999 ,pp. 200–209.
[10] Forman, G., an Experimental Study of Feature Selection Metrics for Text Categorization. Journal of Machine Learning Research, 3 2003, pp. 1289-1305.
[11] Brank J., Grobelnik M., Milic-Frayling N., Mladenic D., "Interaction of Feature Selection Methods and Linear Classification Models", Proc. of the 19th International Conference on Machine Learning, Australia, 2002.
[12] Torkkola K., "Discriminative Features for Text Document Classification", Proc.International Conference on Pattern Recognition, Canada, 2002
[13] Forman, G., An Experimental Study of Feature Selection Metrics for Text Categorization.Journal of Machine Learning Research, 3 2003, pp. 1289-1305.
[14] Sousa P., Pimentao J. P., Santos B. R. and Moura-Pires F., "Feature Selection Algorithms to Improve Documents Classification Performance", LNAI 2663, 2003, pp. 288-296.
[15] Soucy P. and Mineau G., "Feature Selection Strategies for Text Categorization", AI 2003, LNAI 2671, 2003, pp. 505-509.
[16] Kennt Ward Church and Patrick Hanks. Word association norms, mutual information andlexicography. In proceedings of ACL 27, pages 76-83, Vancouver, Canada, 1989.
[17] J.W.Wilbur and k.sirotkin, "The automatic identification of stop words", 1992, pp. 45-55.
[18] Larkey, L.S., A patent search and classification system. Proceedings of DL- 99, 4th ACM Conference on Digital Libraries, eds. E.A. Fox & N. Rowe, ACM Press, New York, US: Berkeley, US, 1999,pp. 179–187.
[19] Amati, G., D'Aloisi, D., Giannini, V. & Ubaldini, F., A framework for filtering news and managing distributed data. Journal of Universal Computer Science, 3(8), 1997 ,pp. 1007–1021.
[20] Weiss, S.M., Apt´e, C., Damerau, F.J., Johnson, D.E., Oles, F.J., Goetz, T. & Hampp, T., Maximizing text-mining performance. IEEE Intelligent Systems, 14(4), pp. 63–69, 1999.
[21] Escudero, G., M`arquez, L. & Rigau, G., Boosting applied to word sense disambiguation. Proceedings of ECML-00, 11th European Conference on Machine Learning, eds. R.L.D. M´antaras & E. Plaza, Springer Verlag, Heidelberg, DE: Barcelona, ES, pp. 129–141, 2000. Published in the "Lecture Notes in Computer Science" series, number 1810.
[22] Drucker,H.,Vapnik,V.&Wu,D.,Support vector machines for spam categorization. IEEE Transactions on neural networks, 10(5),pp.1048-1054,1999.
[23] Drucker, H., Vapnik, V. & Wu, D., Support vector machines for spam categorization. IEEE Transactions on Neural Networks, 10(5), 1999,pp.1048–1054.
[24] Giorgetti, D. & Sebastiani, F., Automating survey coding by multiclass text categorization techniques. Journal of the American Society for Information Science and Technology, 54(12),pp. 1269–1277, 2003.