Contents lists available at GrowingScience

Management Science Letters

homepage: www.GrowingScience.com/msl

Document features selection using background knowledge and word clustering technique

Hajar Farahmand[a*], Ali Harounabadi[b] and S. Javad Mirabedini[b]

[a] *Department of computer engineering, Science and Research Branch, Islamic Azad University, Bushehr, Iran*
[b] *Department of computer engineering, Islamic Azad University, Central Tehran branch, Iran*

C H R O N I C L E

Article history:
Received June 28, 2013
Received in revised format
19 October 2013
Accepted 20 December 2013
Available online
December 29 2013

Keywords:
Background knowledge
Feature selection
Ontology
Word clustering

A B S T R A C T

By everyday development of storage and communicational and electronic media, there are significant amount of information being collected and stored in different forms such as electronic documents and document databases makes it difficult to process them, properly. To extract knowledge from this large volume of documental data, we require the use of documents organizing and indexing methods. Among these methods, we can consider clustering and classification methods where the objective is to organize documents and to increase the speed of accessing to required information. In most of document clustering methods, the clustering is mostly executed based on word frequency and considering document as a bag of words. In this essay, in order to decrease the number of features and to choose basic document feature, we use background knowledge and word clustering methods. In fact by using WordNet ontology, background knowledge and clustering method, the similar words of documents are clustered and the clusters with the number of words more than threshold are chosen and then their frequency of words is accepted as the effective features of document. The results of this proposed method simulation shows that the documents dimensions are decreased effectively and consequently the performance of documents clustering is increased.

1. Introduction

Nowadays, there are massive amounts of human knowledge saved in terms of electronic documents and since they increase rapidly making it difficult to evaluate required information. There are literally various for grouping and indexing documents and these techniques aim to organize a bag of documents in an attempt to increase the speed of accessing to necessary information. Document classification is one of these techniques (Hotho et al., 2003), which assigns natural language documents into a set of predefined categories called document classification.

*Corresponding author.
E-mail addresses: farahmandh@gmail.com (H. Farahmand)

doi: 10.5267/j.msl.2013.12.033

Document space vector is a popular technique used in most of document display classifications. In this model, each document is displayed in terms of a vector of words. One problem with this model, is that the dimension of feature space is very high, which increases the cost and reduces the performance of classification algorithms (Sebastiani, 2002; Shang et al., 2007). Therefore, classification algorithms need some methods to reduce the data size and to increase the classification performance. We need to have another expression to reduce the number of features and it is the main objective of document classification. In order to decrease the number of features, feature selection methods are used (Guyon & Elisseff, 2003). There are several methods for selecting features such as gain information, mutual information, document frequency and correlation coefficient (Forman, 2003; Zheng & Srihari, 2003).

In this paper, to reduce the dimensions of words and to improve the performance of document classification, feature selection methods based on background knowledge and WordNet ontology and word clustering are used. We also choose the clusters with more than a threshold number of words and select the intra-cluster words as the effective feature of document.

This paper is structured as follows: In the second part, the background, some methods in features selection and basic concepts in ontology are presented. In the third part, the earlier work and previous researchers' methods are introduced. The proposed method is presented in fourth part .The results obtained from investigating the proposed method and its effect on improving the performance of document classification are analyzed in the fifth part and in the last part, we try to discuss and make conclusion.

2. Background of the study

Text mining is the process of extracting patterns, and analyzing the relationship and rules between structured or semi-structured documents and in this process, there are various methods of data mining, machine learning, statistical methods, information retrieval and the natural language documents are used.

2.1 Feature selection

Feature selection is a crucial step in the process of document classification. In this step, the best features are chosen for displaying texts. In this paper, in order to increase the performance of document classification, a method for selecting features is proposed.

Different ways to select document features are classified into two various groups: filtering and wrapper methods (Dash & Liu, 1997). Filtering methods are statistical techniques and are independent of the learning method; while wrapper methods take advantage of learning method as evaluation function. Filtering methods, regardless of the learning techniques, by applying a threshold define the number of words, which must be removed from the set of all the words. These methods have low time complexity, but their accuracy cannot be predicted. Among filtering techniques, we can name document frequency, topic document frequency, information gain, mutual information, CHI, SCHI correlation coefficient and Relief-F methods. In words wrapping methods, words are selected due to their effect on increasing the classification accuracy. These methods make use of classifier as the evaluation function and evaluate the impact of each word in the classification accuracy. These methods have high time complexity as well as a very high accuracy. Among wrapping methods, we can consider forward SFS, backward SBS, DTM, PRESET and RC methods (Kohavi & Sommerfield, 1995). In many cases, a combination of these two methods can provide fertile ground to take advantage of both methods. According to researchers, evaluation information gain method is the best method and the document frequency method is introduced as the simplest method (Yang & Pedersen, 1995). In the following, we define the methods used in this paper.

Test χ^2(CHI) method: This method measures the dependency between the feature of *t* and class *c* using χ^2 distribution with one degree of freedom. This test is used in many researches for selecting features in document classification (Bloehdorn et al., 2005). This method in addition to positive correlation information, make use of negative correlation information in weighting features. The amount of test χ^2 (CHI) can be calculated based on Eq. (1). In this equation, *m* represents the number of texts. $P\,(\bar{t}\,,\bar{c}\,)$ denotes the probability that in a document *X* belongs to the training set, feature *t* does not appear and the document does not belong to the class *c*. $P(t,c)$ denotes the probability that in a document *X* belongs to the training set, the feature *t* appears and the document belongs to the class *c*. $P(\,\bar{t}\,,c)$ denotes the probability that in a document *X* belongs to the training set, feature *t* does not appear, but the document belongs to the class *c*. $P(t,\bar{c})$ denotes the probability that in a document *X* belongs to the training set, feature *t* appears in the text, but the document does not belong to class *c*.

$$CHI(t,c) = \frac{m \times \left[P(t,c) \times P\left(\bar{t}\,,\bar{c}\,\right) – P\left(\,\bar{t}\,,\ c\right) \times P(t,\ \bar{c})\right]^2}{P(t) \times P(c) \times P\left(\bar{t}\right) \times P(\bar{c})} \tag{1}$$

Information gain method: In this method, in case we have *ci* class, the obtained information gains from the word *t* shown by *IG (t)* function based on Eq. (2). In this equation, $P(c_i)$ is the probability of class *ci* occurrence, $P(t)$ the probability of word *t* occurrence in a document, $P(c_i|t)$ the improbability of class *ci* occurrence in condition of word *t* occurrence, $P(\bar{t})$ is the probability of word *t* occurrence in a document and $P(c_i|\bar{t})$ improbability of class *ci* occurrence in condition of word *t* occurrence. For each bag of document, information gain is calculated for each unique word and the words with the information gain above threshold are selected.

$$IG(t) = -\sum_{i=1}^{m} P(c_i) + \ \log P(c_i) + P(t) \sum_{i=1}^{m} P(c_i|t)\ \log\ P(c_i|t) + P(\bar{t}) \sum_{i=1}^{m} P(c_i|\bar{t})\ \log\ P(c_i|\bar{t}) \tag{2}$$

Document frequency method: In this method, the number of documents in which the word appears is considered as document frequency for each word (Yang & Pedersen, 1995). For each unique word in the training set, the document frequency is calculated and then the words with a document frequency of less than threshold will be deleted.

2.2 *Ontology*

Ontology is a system of notation symbols that is defined by the symbol of *O = {L, F, C, H, and ROOT}* and includes the following sections (Bloehdorn et al., 2005):

A) *L* is a lexical consists of bag of words.
B) *F* reference function that maps one or more words from the lexical into some equal concept (*F: 2L → 2C*).
C) *C* is a set of concepts.
D) Hierarchical *H*, concepts in the ontology is classified by a relationship-oriented, without cycle, and in the form of transitive and reflection. For example, *H (APPLE, FRUIT)* means the concept of *APPLE* is a sub-concept of *FRUIT*.
E) A high level concept called root (*Root L*) so that for each concept of (*C ε L*) the relation *H (C, ROOT)* is satisfied.

2.3 Strategies of using the concept of the word

There are different strategies for replacing or adding (Hotho et al., 2003):

Add strategy: in this strategy, all of the concepts of each word are added to the word vector of that document in ontology. For example, with the word APPLE the word FRUIT is added to the words vector.

Replace strategy: in this strategy, the equivalent concepts of each word is replaced with each word in document vector, but those words that don't have equivalent concept in ontology will not be deleted from the vector. For example, the word APPLE is replaced with FRUIT.

Only strategy: this strategy is similar to replace strategy except that for those words that don't have equivalent concept in ontology will be deleted from the document vector.

2.4 Disambiguation strategies

Many words have multiple meanings in different documents and inserting all the equivalent meaning of each word to document vector, besides increasing the dimensions of document vector, reduce the classification accuracy. To resolve this ambiguity, there are different strategies:

All strategy: in this strategy all the concepts associated with each word are added to the document vector. Thus, this strategy does not eliminate ambiguity.

First strategy: usually in ontologies such as Word Net concepts of each word are arranged in the order of the common meaning of that word in the language. In this strategy, the first meaning of the word in ontology is the most likely meaning.

Context Strategy: this strategy tries to have a suitable map of each word to its equivalent concept according to the document content.

2.5 *Strategy of hierarchies of concepts*

The main objective of this strategy is adding sub-concept frequency in a document to its higher level concept in the hierarchical of ontology concepts. Frequency of each concept with the total frequency of r its next sub-concept is updated in the ontology hierarchical. If R = 0 frequency of each concept is independent of the frequency of its sub-concept and only returns the concepts which refer directly to the word. $R = n$ means the frequency of each concept with the frequency of n level is gathered from its lower level concepts and $R = \infty$, i.e., the frequency of each concept with the frequency of all of its sub- concepts are gathered in the ontology. The researchers evaluation by considering five level of sub- concepts (R =5), the best results have been found.

3. Related works

Many studies have been accomplished on document classification and in some of these studies the focus has been on the different techniques of word choice, the way of weighting, and the document representation space. In addition, some of the others put into consideration the various methods and algorithms in classification, methods in learning machine or statistical methods and their effect on the performance of document classification.

Forman (2003) for classification of English documents made use of 12 different methods that among such methods we can mention information gain, document frequency, CHI, BNS and Odd Ratio methods. In the document frequency method for each unique word he calculated the document training set in which the words appears as the document frequency and then delete the words with the document frequency of less than threshold. The underlying assumption of this method is basically that the words that are less useful for predicting the class that the document belongs to are not appropriate or are not effective in the overall performance. Anyway deleting the words with less usefulness will decrease the words dimensions and if some words with less document frequency be Words of miss (error) and noise their elimination will also increase the classification accuracy.

Forman also measured the amount of dependency between term *t* and class *c* in the test χ^2(CHI) method. In this method, he also made use of the negative correlation of the information in addition to positive correlation of the information weighting of a word. He calculated this amount for each word of *t* in each class of *c* and finally, he considered the maximum values as the CHI criteria of that word and selected the highest of them.

Kononenko (1994) used a filtering method called Relief-F for the choice of words in the document domain that had good results in the deletion of words for document classification. In this method, he used two concepts of nearest hit and nearest miss in selecting word i.e. in each document the values of the nearest hit and the nearest miss of that document and the Euclidean distance among that document and the found documents was calculated and then the average of this amount was used in order to update words weight. He eventually selected the words with the most weight among the weighted words.

Yang and Pedersen (1995) studied five different ways of selecting word: information gain, mutual information, document frequency, CHI and the power of words on a set of standard data Reuters and OHSUMED using the algorithm of KNN and LLSF. In fact, they made use of filtering methods that are part of the statistical methods in these methods; words are selected by applying a threshold. In the information gain for each unique word by using equation (2), that was previously described, and the words with the information gain of less than a predetermined threshold were removed from the feature space. In addition, in the method of mutual information, they calculated the amount of dependency for each word of t in each class of c. Finally, they selected the most amount of it, as the mutual information of that word and the words with the amount of mutual information above threshold were selected.

Word clustering is also an effective method to reduce feature dimensions and distribution and, as a consequence, improves the performance of document classification (Han et al., 2005) which is becoming a key technique in natural language processing issues. They introduced a method for clustering words based on domain rules and syntactic structure. In their method, clusters were formed from databases with different domains and different orthographic features (Schone & Jurafsky, 2001). In fact, the words clustering were based on a priori knowledge of a particular class. For instance, the words “Merry”, “Johnson” and “Tom” in the database are clustered in the database of name word and similarly the word “Massachusetts” is clustered in the database of the word situation. In their methods orthographic features of word included examples of words, numbers and special symbols of words. For example, “@” is an orthographic feature of e-mail address used in clustering e-mail addresses in document.

According to Han et al. (2005), words clustering have been developed based on the rules in three steps include creating a database of domains, clusters designing and rules designing. As a results, while reducing the dimension, a full recovery in 6.6 percent have been on the average of classification performance of document header lines and an improvement of 4.8% on the overall accuracy of extracting bibliographic fields. In addition, for extraction of document metadata the given method by them has had better results than words distributional clustering method.

4. The proposed method

The process of classifying documents in the proposed method as shown in Fig. 1 is carried out generally, in five stages: document pre-processing, stop word removing and stemming, indexing and weighting documents vectors, feature selection by using words clustering and background knowledge, exerting the support vector machine algorithm and evaluation.

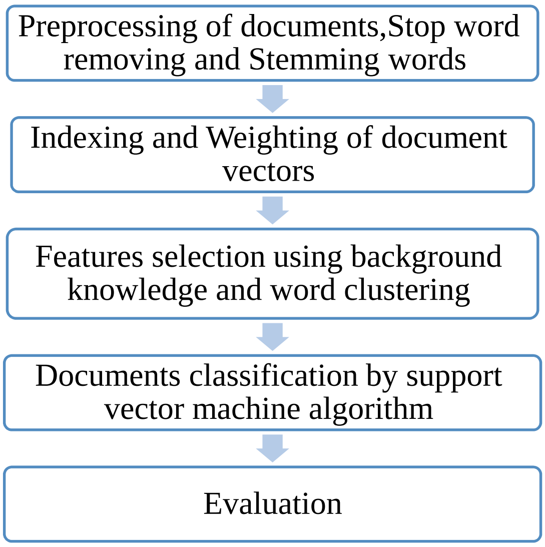
Preprocessing of documents,Stop word removing and Stemming words
Indexing and Weighting of document vectors
Features selection using background knowledge and word clustering
Documents classification by support vector machine algorithm
Evaluation

Fig.1. Process of classifying documents in the proposed method

4.1 Preprocessing of documents

The first step in the preprocessing of documents is preprocessing. This step involves stop word removing, eliminating tags, and stemming words (Breaux & Reed, 2005). Performing this step helps to reduce the space of many words. Common words refer to some useless words that from the linguistic point of view are of no content value in document, and they can be find in all of the documents with high frequency such as prepositions, condition, connection, pronouns, etc. These words have no effect in distinguishing a document from other documents and deleting them has a considerable impact on increasing computational speed and reducing the number of words (Makrehchi & Kamel, 2004). To do this, we can use a fix list of a bag of common word and by comparing it with document ,identify and delete the common words or by POS tagging, determine the syntax of each word in a sentence and then delete all the words that are not nouns, verbs or adjectives. XML, HTML tags, punctuation marks, special signals as common words are useless as common words and should be deleted. In this paper in order to delete common words, the fixed method is used. The next step in the processing of documents is stemming words. In fact stemming aims to delete prefixes and suffixes of words and finding the root of each word. This process uses a series of rules; each word is converted to its root and the proposed model of this paper, the Porter algorithm is applied.

4.2 Documents indexing

After preprocessing stage, each document is represented as a bag of words. In order to apply document clustering methods and to apply document word selection methods, a good structure should be chosen for displaying documents. The most common method of representing documents is vector representation. In this method, the document is considered as a vector of words. At this point, the repeated words of each document are identified and their frequency is calculated. Then the weight of each word in the document is determined and finally the bag (set) of documents is represented as a matrix or table. Table 1 shows the document representing method as a vector.

Table 1
The document representing method as a vector

Documents	Words Space			
	F_1	F_2	...	F_m
D_1	W_{11}	W_{22}	...	W_{1m}
D_2	W_{21}	W_{22}	...	W_{2m}
...	...	...	...	...
D_n	W_{n1}	W_{n2}	...	W_{nm}

As Table 1 shows the set of $F = \{F_1, F_2 ... F_m\}$ denotes the words space and the set of $D = \{D_1, D_2 ... D_n\}$ represents the set documents. Each value of W_{ij} represents the weight of the word f_j in the document D_i.

There are different weight structures for document indexing. One of the best and most popular methods in weighting words is *TF × IDF* weighting method that is used in this paper. In this method, the weight of each word in the document is obtained by multiplying the frequency of the word in the inverter frequency of the document. Therefore, the repetition of a word in a document is effective in weight gain of it only when it is not repeated in other documents and is usually defined in the form of Eq. (3) (Sebastiani, 2002).

$$\mathrm{W_{jk}} = \mathrm{F_{jk}} \times \log\frac{N}{\mathrm{N(F_j)}} \tag{3}$$

In Eq. (3), F_{jk} is the frequency of the word j in the document K, N is the total number of documents, $N (F_j)$ is the number of documents that the word F_j occurred in it at least once.

4.3 Features selection

Too much number of features is a major obstacle for many machine learning techniques. If it is a small number of selected features, accuracy and performance of the classification algorithm will reduce, in contrast the large number of features also increases the time complexity of the classification algorithms as a result decrease performance of classification. Therefore reducing the feature space, without loss of accuracy of classification can reduce the time complexity of document classification algorithms and consequently is considered as one of the main aims of document classification.

As described in Section 2.1 in the classification process, we often make use of statistical methods and the threshold for selecting features and the words conceptual similarities and semantic relation does not have a role in the selection of features (Hotho et al., 2003) while background knowledge and ontologies can be used for semantic similarity of words (Bracewell et al., 2005; Fellbaum, 1998). Using ontology in addition to solving the problem of synonymous words provides taking the advantage of the general concepts and higher level concepts in the hierarchy of concepts. For instance, methods of word choice cannot find any relationship between the two words ORANGE AND APPLE while by using the high-level concept of FRUIT instead of the two words ORANGE AND APPLE in document vector, the semantic relationship between the two texts is easily recognizable. This paper uses conceptual similarity and semantic relation of words and words clustering in each class of the training documents for selecting words, Therefore, by combining semantic concepts and by ontology with statistical method based on the frequency of clusters in selecting word, has benefited from the advantages of both methods.

In this paper, the conceptual similarity and semantic relation of words are used according to background knowledge and WordNet ontology in selecting words. There are various strategies for replacing or adding concepts instead of words that we made use of the only strategy in order to replace words. Many words have different meanings in different contexts and inserting all the equivalent meanings of each word into the document vector may result in increasing vector dimensions. In addition, to decrease the performance of document classification to resolve this ambiguity there are a variety of strategies; in this paper the first strategy, has been used to resolve this ambiguity and also for using the concept of hierarchy of concepts the hierarchical strategy has been used to five levels. One of the most popular clustering algorithms is k-means and in this paper, we used it for clustering. The similarity criteria in this algorithm is Euclidean distance between points that in this paper since we used wordnet ontology the defined semantic identification is used for each word in the word net tree as the space of each word in the vector space.

In this paper, the words of training documents belonging to each category to choose its best feature is clustered by K-means method and the clusters with the number of words higher than threshold are selected and the words within them are selected as the main features of that class.

4.4 Applying support vector machine algorithm

According to the results of the previous researches, the classifier of support vector machine is considered as one of the best classifier of the documents and in this paper after choosing the words of each class we have done the classification process by support vector machine. One of the benefits of this method is that it is not dependent on the number of training samples and with a high number of features and small number of samples can also act well. The approach of this method is in a way that tries to choose the decision boundary in the training phase in a way that the minimum distance of it with any of the desired classes be the maximum.

4.5 Evaluation

After operating the above four steps, the methods of gain information, document frequency and CHI test that used statistical methods in feature selection have been used to compare and evaluate the results of the proposed methods that the results are mentioned in the next section.

5. Results

To evaluate the proposed method, a subset of the Reuters 21578 data set is used (Lewis et al., 2004). Reuters 21578 is a standard data set that is used in many studies to be associated with text mining. Subset of the data used in this paper contains 3647 documents that 2666 of them are train documents and the other 981 are used as the test documents. This subset consists of eight classes of the main ten classes of Reuters 21578 dataset. The information of this subset is shown in Table 2.

Table 2
Subset of the Reuters 21578 data set

Class label	Total number of documents	Number of train documents	Number of test documents
money-fx	717	538	179
grain	582	433	149
crude	578	389	189
trade	486	369	117
interest	478	347	131
ship	286	197	89
wheat	283	212	71
corn	237	181	56
Total	3647	2666	981

In this paper, in order to evaluate the performance of classification algorithms, we made use of macro F-measure and micro F-measure. In macro F-measure, the value of *F* is calculated for each class and then the average of all classes is calculated. Thus to each class, regardless of its frequency, equal weight is attributed while in the micro F-measure the value of F is calculated generally, and without distinguishing between classes and for the entire data set. The criteria are described by Eq. (4-7),

$$\text{MacroF1} = \frac{\sum_{k=1}^{C} F_k}{C},\ F_k = \frac{2P_k \times R_k}{P_k + R_k} \tag{4}$$

P_k and R_k in the macro F-measure equation are calculated by Eq. (5) and Eq. (6) relations.

$$\text{Precision } : P_k = \frac{TP_k}{TP_k + FP_k} \tag{5}$$

$$\text{Recall } : R_k = \frac{TP_k}{TP_k + FN_k} \tag{6}$$

$$\text{MicroF1} = \frac{2P \times R}{P + R} \quad (7)$$

TP_k is the number of documents correctly classified in the class C_k by the algorithm.

FP_k is number of documents that belonged to other classes and have been wrongly attributed to the class C_k.

FP_k is the number of texts that belong to class C_k, but the classification algorithms by mistake classified them in different classes (Sebastiani, 2002).

P is the precision and R is the recall that are calculated on the entire data set their values are calculated similar to Eq. (5) and Eq. (6) to formulas with the difference that their values are not for every class, but are taken for the entire data set. Results obtained from macro F-measure on the mentioned data set are shown in Table 3 and the results of micro F-measure in Table 4.

Table 3
Results obtained from macro F-measure

Number of features	Feature selection methods			
	CHI	DF	IG	Proposed method
25	0.50	0.38	0.45	0.59
50	0.56	0.42	0.55	0.64
100	0.66	0.51	0.65	0.72
200	0.70	0.57	0.70	0.78
400	0.76	0.65	0.75	0.81
800	0.76	0.70	0.76	0.84
1000	0.76	0.73	0.76	0.84

As we can observe from the results of Table 3, the value of F is calculated for each class and then the average of all of the classes are calculated so that the proposed method of this paper selects the words of each class by word clustering, separately and presents better results for the number of various words in macro F-measure. The IG and CHI methods have approximately the same and similar results and the worst performance is belongs to the DF methods.

Table 4
Results obtained from micro F-measure

Number of features	**Feature selection methods**			
	CHI	DF	IG	**Proposed method**
25	0.61	0.53	0.60	**0.60**
50	0.75	0.59	0.74	**0.71**
100	0.76	0.64	0.75	**0.73**
200	0.80	0.68	0.78	**0.81**
400	0.82	0.73	0.81	**0.82**
800	0.83	0.78	0.83	**0.84**
1000	0.84	0.81	0.84	**0.85**

In the micro F-measure that the result of it is shown in the Table 4 is a proposed method that has the same or better results than the other methods for the number of words more than 100. CHI methods as statistical methods have better results than the other statistical methods however the IG method had close and similar result. The same as in macro F-measure the DF method has the lowest results even in this criteria.

6. Conclusion

One of the important issues in document classification is the high dimension of words in representing vector space. Lots of words in the documents are redundant and irrelevant and result in reducing the performance of classification algorithms. In this paper, after preprocessing the documents, the effective method of TF-IDF has been used. In order to select effective words, the word net background knowledge and words clustering methods are used and the obtained results from document classification according to proposed method has been compared with word selection, information gain, CHI, document frequency methods on a subset of the Reuters 21578 data set. The results of evaluation show the effectiveness of the proposed method especially in macro F-measure.

References

Bloehdorn, S., Cimiano, P., Hotho, A., & Staab, S. (2005, May). An Ontology-based Framework for Text Mining. In *LDV Forum*, 20(1), 87-112.

Bracewell, D., Ren, F., & Kuroiwa, S. (2005). Multilingual Single Document Keyword Extraction for Information Retrieval. In *Proc. IEEE Int. Conf. Natural Language Processing and Knowledge Eng*, 517-522.

Breaux, T. D., & Reed, J. W. (2005). Using Ontology in Hierarchical Information Clustering. In *Proc. 38th Hawaii Int. Conf. System Sciences.*

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, *1*(3), 131-156.

Fellbaum, C. (1998). WordNet: an electronic lexical database. *MIT Press*.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, *3*, 1289-1305.

Guyon, I., & Elisseff, A. (2003). An Introduction to variable and feature selection. *Machine Learning Research* 3, 1157-1182.

Han, H., Manavoglu, E., Zha, H., Tsioutsiouliklis, K., Lee Giles, C., & Zhang, X. (2005). Rule-based Word Clustering for Document Metadata Extraction. *ACM Symp. On Applied Computing,* 1049-1053).

Hotho, A., Staab, S., & Stumme, G. (2003). Ontologies Improve Text Clustering. In *Proc. ICDM'03 3rd IEEE Int. Conf. on Data Mining* (pp. 541).

Kohavi, R., & Sommerfield, D. (1995). Feature subset selection using wrapper methods: verfitting and dynamic search space topology. In *Proc. 1st International Conference on Knowledge Discovery and Data Mining* (pp. 192-197).

Kononenko, I. (1994). Estimating attributes: analysis and extension of RELLIEF. In *Proc. 6th European Conference on Machine Learning* (ECML-94) (pp. 171-182).

Lewis, D.D., Yang, Y., Rose, T., & Li, F. (2004). RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research,* 5, 361-397).

Makrehchi, M., & Kamel, M. (2004, June). A fuzzy set approach to extracting keywords from abstracts. In *Fuzzy Information, 2004. Processing NAFIPS'04. IEEE Annual Meeting of the* (Vol. 2, pp. 528-532). IEEE.

Schone, P., & Jurafsky, D. (2001). Knowledge-free induction of inflectional morphologies. In *Proc. North American Chapter of the Association for Computational Linguistics on Language technologies* (pp. 1-9).

Sebastiani, F. (2002). Machine learning automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.

Shang, W., Huang, H., Zhu, H., Lin, Y., Qu, Y., & Wang, Z. (2007). A novel feature selection algorithm for text categorization. *Expert Systems with Applications,* 33(1), 1–5.

Yang, Y., & Pedersen, J.P. (1995). A comparative study on feature selection in text categorization. *14th Int. Conf. Machine Learning* (pp. 412–420).

Zheng, Z., & Srihari, R. (2003). Optimally combining positive and negative features for text categorization. *ICML Workshop*.