DOCUMENT CLASSIFICATIONS BASED ON WORD SEMANTIC HIERARCHIES

Xiaogang Peng & Ben Choi

Computer Science, College of Engineering and Science
Louisiana Tech University, LA 71272, USA

pro@BenChoi.org

Abstract

In this paper we proposed to automatically classify documents based on the meanings of words and the relationships between groups of meanings or concepts. Our proposed classification algorithm builds on the word structures provided by WordNet, which not only arranges words into groups of synonyms, called Synsets, but also arranges the Synsets into hierarchies representing the relationships between concepts. Most existing methods classify text documents based on the number of occurrences of words and some based on Synsets. Our approach goes one step further by using not only word occurrences and Synsets but also the relationships between Synsets. We also proposed a sense-based document representation based on the semantic hierarchies provided by WordNet. To classify a document, our approach extracts words occurred in the document and uses them to increase the weight of the Synsets corresponding to the words. Words with same meanings will increase the weight of their corresponding Synsets. As a result, we count the occurrences of senses. We also propagate the weight of a Synset upward to its related Synsets in the hierarchies and thus capture the relationships between concepts. In comparing to previous research, our approach increases the classification accuracy by 14%.

Keywords: Classification, WordNet, Semantic Web, Document Representation, Information Retrieval

1. Introduction

Automatic text classification is the task of assigning a text document to a relevant category or categories. Formally, let C = $\{c_1, \ldots, c_K\}$ be a set of predefined categories, D = $\{d_1, \ldots, d_N\}$ be a set of text document to be classified. The task of text document classification is then transformed to approximate the unknown assignment function *f*, which maps $D \times C$ to a set of real numbers. Each number in the set is a measure of the similarity between a document and a category. Based on the measures, a document is assigned to the most relevant categories [4].

Document representation is one of the most important issues in text classification. In order to be classified, each document should be turned into a machine comprehendible format. The *bag-of-words* document representation [11, 13] is simple, yet limited. Attempts have been conducted to improve the effectiveness of the representation. For example, Mladenic [16] extends the "bag-of-words" to the "bag-of-phrases" and showed improvement of the classification results [2].

There are two major problems with the bag-of-words or the bag-of-phrases representations. First, it counts word occurrences and omits the fact that a word may have different meanings (or senses) in different documents or even in the same document. For example, the word "*bank"* may have at least two different senses, as in the "*Bank"* of America or the "*bank"* of Mississippi river. However counting word occurrences, these two instances of "*bank"* are treated as a same feature. The second major problem lies in the fact that sometime related documents may not share the same keywords so that two related documents cannot be recognized as belonging to the same category. Thus, rather than counting word occurrences, counting word senses might improve text classification. Sense based text classifications [19, 13] are attempts to address the problems.

However, we discovered that the previous sense based document classifications [1, 7, 19,] did not make use of semantic hierarchy of senses. We proposed that after word senses are extracted from a document, all the senses should be considered globally, from the point of view of the entire document, instead of treating each sense separately. We considered that the most widely used "bag of words" or "bag of senses" representations of a document are not sufficient to represent the global relationships of senses. Making use of the relationships between word senses provided by the hierarchical structures of Synsets in WordNet [15], we proposed a new document representation that exploits the semantic hierarchy and developed a corresponding semantic hierarchy classification system.

This research was funded in part by a grant from the Center for Entrepreneurship and Information Technology (CEnIT), Louisiana Tech University.

Xiaogang Peng & Ben Choi, "Document Classifications Based On Word Semantic Hierarchies,"
The IASTED International Conference on Artificial Intelligence and Applications, pp.362-367, 2005.

The remainder of this paper is structured as follows. Related research is provided in Section 2. A new document representation based on word semantic hierarchies is defined in Section 3. A new document classification system based on the new document representation is described in Section 4. Testing and performance analysis of the new classification system is provided in Section 5. And, the conclusion and future research are provided in Section 6.

2. Related Research

The commonly used text document representation is the "bag-of-words", which simply uses a set of words and the number of occurrences of the words in a document to represent the document [11, 13]. Many efforts have been taken to improve this simple and limited document representation. For example, Mladenic [16] uses phrases or word sequences to replace single words, for which Chan [2] confirmed the improvement of the approach by experiments. The goal of using phrases as features is to attempt to preserve the information left out by the "bag of words" methods. This results in a document representation called "feature vector representation" that uses a feature vector to capture the characteristics of a document by an "N-gram" feature selection. An N-gram feature could be a word or a sequence of N words. Experiments showed that N ranging from two to three is sufficient in most classification systems.

Since the number of different words and the number of two to three sequence of words in a document can be very large that in turn results in large computational cost, various techniques have been employed to reduce the number of features. The most frequently used methods to reduce the number of features are "stopping" and "stemming". The idea of "stopping" is to eliminate those common words that occur often and mean little, such as articles or prepositions. The "stemming" on the other hand is trying to use a language-specific stemming algorithm to find the same semantic root of different words, such as "compute" and "computes" are considered as the same feature.

For text document classification, TFIDF (Term Frequency–Inverse Document Frequency) method is often used. It represents each document as a "TFIDF" vector in the space of features (word or phase) that are taken from training documents, then sums up all the document vectors and uses the resulting vector as a model for classification. The term frequency $TF(f_i, Doc)$ of a feature f_i in a document *Doc* is calculated by counting the number of occurrences of f_i. Let T be the total number of documents and $DF(f_i)$ be the number of documents having the feature f_i, the inverse document frequency of a feature f_i, denoted by $IDF(f_i)$, is usually defined as:

$$IDF(f_i) = Log\frac{T}{DF(f_i)}$$

A document is represented by a vector with each item i defined as:

$$V(i) = TF(f_i, Doc)IDF(f_i).$$

The TFIDF is extended by Joachimes [8] who analyzed the TFIDF classifier in a probabilistic way based on the implicit assumption that the TFIDF classifier is as explicit as the Naïve Bayes classifier. By combining the probabilistic technique from statistic pattern recognition into the simple TFIDF classifier, he proposed a new classifier called the PrTFIDF classifier. The PrTFIDF classifier optimizes the parameter selection in TFIDF and reduces the error rate in five out of six reported experiments by 40%.

Other more sophisticated machine learning methods and classification algorithms can be applied to induce representations for categories from the representations of documents. A text classification system, that takes advantage of the hieratical structure of categories, is reported by Choi and Peng [3]. Other related classification methods can also be found in [4].

To move from counting word occurrences to counting senses, a database of senses is required. We choose WordNet [14, 15] as the database to help the process of document representation and classification. The basic unit in WordNet is called synonym set or Synset. Each Synset consists of a list of synonymous word forms. A word form in WordNet can be a single word or two or more words connected by underscores. WordNet is capable of referring a word form to a Synset. All the Synsets are divided into five categories: Nouns, Verbs, Adjectives, Adverbs, and Function verbs. In each category, the Synsets are organized by semantic relations, some of which are listed as follows:

- Hyponym / Hypernym: The "is-a" semantic relation or subset/superset relation. Hyponymy is transitive and asymmetrical. For example, *economic* is a hyponymy of *social science,* but *social science* is a hypernym of *economic.*
- Meronym / Holonym: The **"**has-a" relation. If the sentence "An *x* is a part of *y*" is meaningful, then x is the meronym of y and y is the holonym of x.

The relation that interests us here is the hyponym/ hypernym relation between nouns. A Synset is a hypernym of another if it covers a more general meaning. For example, `science` is a hypernym of `natural science` and `social science` since it represents a more general concept. Based on the relations, Synsets in WordNet are organized into tree structures.

WordNet is widely used for sense-based projects. For instance, Rodriguez [18] used the synonymy in WordNet and showed an improvement in classification accuracy on the Reaters-21578 corpus. Scott and Matwin [19] used synonymy and hypernymy to develop a "*hypernym density representation"* and achieved a small improvement in the classification accuracy.

3. Semantic Hierarchy Representation

The first issue that needs to be addressed in document classification is how to represent a document so as to facilitate machine manipulation but also to retain as much information as needed. If senses are used to represent a document, the relations between senses play a key role in capturing the ideas in the document. Recent research shows that simply changing the keywords to senses without considering the relations does not have a significant improvement and sometime even perform worse than keywords [10]. In an attempt to address the problem, different semantic information is incorporated, such as Scott and Matwin [19] used the "is-a" relationships and showed a minor improvement.

We proposed a method to account for the semantic relations. We call our document representation as semantic hierarchy representation. The basic idea of our representation is to represent a document by a group of hierarchical senses, which makes use of the organization of senses provided by WordNet.

To generate a document representation using our proposed method, two steps are required:

(1) Mapping words into Synsets, and

(2) Capturing relationships between Synsets

In step (1), we extract words occurred in a document and use them to increase the weight of the Synsets corresponding to the words. Words that have the same meanings will increase the weight of their corresponding Synsets. As a result, we are counting the occurrences of senses or meanings. In step (2), we propagate the weight of a Synset upward to its related Synsets in the hierarchies of WordNet and thus capture of the relationships between senses or concepts. Details of these two steps are provided as follows.

3.1 Mapping Words into Synsets

The process for mapping words into Synsets is illustrated with an example shown in Figure 1. For simplicity, supposed there is a document consisting only 10 distant words or word phrases: government (2), politics (1), economy (1), natural philosophy (2), life science (1), math (1), political economy (1), and science (1), where the number indicated the number of occurrences, which is turned into percentage and shown in Figure 1. The words or word phrases are then mapped into their corresponding Synsets in WordNet. In the example, government (0.2) and politics (0.1) are mapped into the Synset government, and the weight 0.2 and 0.1 are added to 0.3 as indicated in Figure 1.

The above described an oversimplified example where there is a one-to-one mapping from one word to one Synset. However, one word may have several meanings and thus one word may be mapped into several Synsets. In this case, we need to determine which meaning is being used, which is the problem of sense disambiguation [22]. Since a sophisticated solution for sense disambiguation is often impractical, we propose a naïve approach that consists of the following four stages:

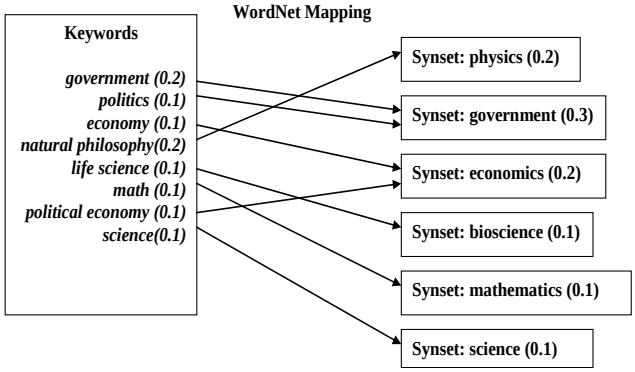


Figure 1 Example of mapping keywords into Synsets

(1) A word is simply mapped into a Synset or Synsets that contain the word.

(2) We increase the weights of the mapped Synsets and all their hypernyms and hyponyms by one.

(3) We process all the words in the document repeating the stages (1) and (2).

(4) We select the most relevant Synset for the word. We consider the Synset that has the highest weight being the most relevant. If all the Synsets of a word have equal weight, then we select the Synset that represents the most often used sense of the word.

This naïve approach is based on the assumption that a document contains central themes that in our cases will be indicated by certain Synsets having height weights.

3.2 Capturing Relationships between Synsets

To capture the relationships between Synsets, we propagate the weight obtained in step 1 (Section 3.1) for each Synset from leaf nodes to the root following the tree hierarchies provided in WordNet. Figure 2 gives a visual example of this step. The six Synsets with different original weight will be propagated in the directions as shown by the arrows in the figure. For instance, the weight for bioscience and that for Physics are propagated up to contribute to the weight of Natural Science.

As the weights are contributed to the upper nodes, they are scaled and added to the weight of the upper nodes. For the scaling, we utilize Formula 1 and 2, for which we consider the WordNet noun Synset hierarchy as a tree T by taking each Synset as a tree node. A subtree T_N, whose root node is node N, has k children nodes labeled from N_1 to N_k. As a special case, when k=0, T_N is a *leaf node*. T_N has k *direct subtrees*, whose root nodes are N_1 to N_k

$$W'(T_N) = W(N)\alpha(N,T_N) + \sum_{i=1}^{k} W'(Sub_N(N_i))\beta(Sub_N(N_i),T_N)$$

Formula 1 Formula for Calculating Propagated Weight

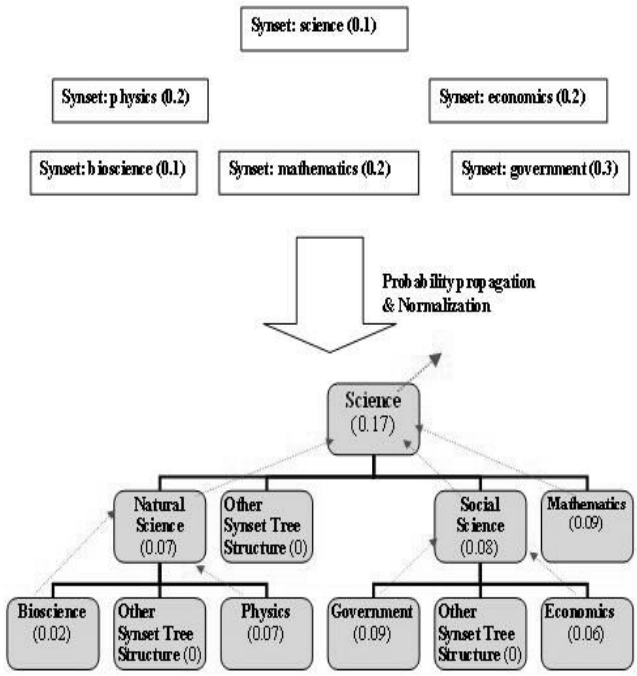


Figure 2 Turning Synsets to Semantic Hierarchy

correspondingly. $Sub_N(N_i)$ is a direct subtree rooted at child node N_i. The original weight $W(N)$ of a node N is the weight obtained in step 1 (Section 3.1).

The propagated weight $W'(T_N)$ for tree T_N is calculated using Formula 1 which is devised based on [16]. Formula 1 is used recursively starting from the root N and stops in the leaf nodes. In Formula 1, the propagated weight of the tree T_N is composed by the original weight of the root node multiplied by a scale factor $\alpha(N, T_N)$ and the propagated weight of each direct subtree multiplied by the corresponding scale factor $\beta(Sub_N(N_i), T_N)$. These two scale factors are defined in Formula 2 (based on [16]), which is obtained by using the size of node N and the size of each direct subtree ($Size(Sub_N(N_i))$). The size of node N is defined as the number of synonyms within the Synset corresponding to node N. Similarly, $size(Sub_N(N_i))$ is the number of synonyms within all the Synsets in the subtree $Sub_N(N_i)$.

$$\alpha(N, T_N) = \frac{\ln(1 + Size(N))}{\ln(1 + Size(N)) + \sum_{i=1}^{k} \ln(1 + Size(Sub_N(N_i)))}$$

$$\beta(Sub_N(N_i), T_N) = \frac{\ln(1 + Size(Sub_N(N_i)))}{\ln(1 + Size(N)) + \sum_{i=1}^{k} \ln(1 + Size(Sub_N(N_i)))}$$

Formula 2 Scale Factors for Calculating Propagated Weights

4. Sense Based Classification System

In the last section, we described our semantic hierarchy document representation. In this section, we extend the representation to define categories. After categories are defined, the task of classification is to assign documents to the categories. In the following, we describe how to define categories and how to classify documents based on the semantic hierarchy representation.

4.1 Defining Categories

In order to define a category, we first need to determine what source we should use as examples of categories. In many existing classification systems, a set of training examples or documents is used to generate the representation of a category. There are two major problems with this method. The first one lies in the fact that the training examples might contain a lot of unrelated information and existing learning methods cannot eliminate the pollution. On the other extreme, if there are only few training examples available, then the representation of the category might not be sufficient.

For our sense based classification system, we proposed to extract category information from the name of the category, keywords from the explanation of the category, and meronym of the keywords provided by WordNet. There are advantages by extracting category information in this way. The name of a category is usually the most informative part for a category thereby reducing the risk of polluting the actual meaning of the category. Also, keywords in the explanation and meronyms of the keywords are additional sources to gain additional information for defining the category. Our experiments provided in Section 5 confirm that using our approach improves the classification accuracies.

Once we obtained a list of keywords as described above, we generate a semantic hierarchy representation of each of the predefined categories by using the steps provided in Section 3.

4.2 Classifying Documents

To classify a document, our system first extracts features (words or word phrases) from the document and use the features to build a semantic hierarchy representation by the methods described in Section 3. Since the predefine categories for our system are also encoded using semantic hierarchy representations, the classification task becomes finding a method to compare two semantic hierarchy representations. The process of classification then is to compare the document representation to each of the categories representations. Then, the document is assigned to the categories having the closest match or similarity.

We defined the similarity between a document d and one of the given categories c_k by Formula 3, for which we let n be the number of Synsets in the WordNet noun database, $c_{k,l}$ and d_l as the propagated weights of the corresponding Synset l in the semantic hierarchy representation of the document and the category, respectively. Then, after checking all the given categories, the document d is classified to the category that has the maximum similarity with the document.

$$Sim\ (d, c_k) = \frac{\sum_{l=1}^{n} (d_l c_{k,l})}{\sqrt{\sum_{l=1}^{n} d_l^2} \sqrt{\sum_{l=0}^{n} c_{k,l}^2}}$$

Formula 3 Similarity of a Document and a Category

5. Testing and Performance Results

In this section, we design experiments to test our proposed methods. The first part is designed for testing the methods for selecting the sources to define categories. The second part is designed to evaluate the performance of our classification system in comparing to other related methods.

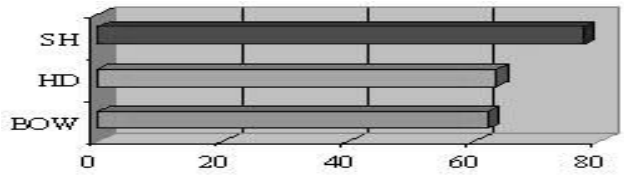
5.1 Experiments on Defining Categories

We designed a set of experiments to compare the effectiveness of our method of defining categories with a related method described in [12]. For the experiments, we use a section of Yahoo.com directory as test base. The predefined category hierarchy is taken from a subtree of Yahoo.com directory. In the experiments, the impact of using two different sources to describe the categories is tested. The first source (as proposed in [12]) is using the keywords of the summaries and titles of the web pages provided in Yahoo categories. The other source is using the name of the category, keywords from the explanation of the category, and meronym of the keywords provided by WordNet as described in Section 4.1.

To compare the two different sources for defining the categories, we use 200 pre-classified web pages taken from a two-level category hierarchy provide by Yahoo.com. The results of the experiments are listed in Table 1. Labels in Table 1 are defined as follow: *Correct* stands for the result that a web page is classified to the category where it is taken. *Not Deep Enough* means a web page is classified to the parent category of its original category. The classification system that we used in the experiments is one that can dynamically add new categories (*Expanded*) and that uses a hierarchical category level structure (*Error in levels*) as reported in [3]. In here we are most concerned with the correct rate. The experiments show that our method

Table 1 Classification Results on Evaluation of the Effectiveness of Defining Categories

	DT	TM
Correct	**34**	**62**
Not Deep Enough	26	76
Expanded	2	0
Error in 2nd level	109	42
Error in 1st Level	29	20



SH	HD	BOW
77.46	63.57	62.14

Figure 3 Comparisons of three Classification Systems

(TM as shown in Table 1) got 62 correct results while the other method (DT) got 34 correct results. This amounts to an 82% improvement by using our method for our classification system. These results reflect the specific need for our semantic hierarchy classification and may not be applicable for other systems.

5.2 Comparing our System with Related Systems

We compare the performance of our semantic hierarchy classification system with two other related systems: hypernym density [19] and the usual bag-of-words system that is used as a baseline. To get a fair comparison, we choose to use the same newsgroups for our experiments as for [19]. The newsgroups are bionet.microbiology and bionet.nueroscience from Usenet, which is challenging for classification due to the posting by different users all around the world, using different terminology and special writing styles. We defined categories using the method described in Section 4.1. Then we randomly select 217 postings, 98 in microbiology and 119 in neuroscience to test the classification systems.

The result of the accuracy rate from three classification methods: our "Semantic Hierarchy" (SH), "Hyperynym Density" (HD), and "Bag of Words" (BOW) are summarized in Figure 3, which shows that our method (SH) has the highest accuracy rate at 77.46% while the other two systems only achieves accuracy rates at around 63%. This shown that our method has a 14% improvement in the classification accuracy. Again, these results may be domain specific, in this case for newsgroups, and the results for other domains may vary.

6. Conclusion and Future Research

In this paper we proposed to automatically classify text documents based on the meanings of words and the relationships between groups of meanings or concepts. We proposed a semantic hierarchy representation and a corresponding classification system. To classify a document, our approach extracts words occurred in the document and uses them to increase the weight of the Synsets corresponding to the words. Words that have the same meanings will increase the weight of their corresponding Synsets. As a result, we are counting the

occurrences of senses or meanings. We also propagate the weight of a Synset upward to its related Synsets in the hierarchies and thus make use of the relationships between concepts. In comparing to a previous research, our approach increases the classification accuracy by 14%.

We also experimented on selecting different sources to define categories and found that it has signification effect on the overall accuracy of the classification system. In particular, our method of using the name of the category, keywords from the explanation of the category, and meronym of the keywords provided by WordNet out performs a related method of using the keywords of the summaries and titles of the web pages provided in Yahoo categories by 82%. However, we should point out that this result reflects the specific need for our sense based classification and may not be applicable for other systems that do not take advantage of word senses.

Our work shows promising future of applying semantics for classifications. It also shows that relationships between groups of meanings or concepts are promising sources for mining semantic information from documents. Much work can be done in this direction on the move to the future of semantic information age.

References

[1] G. M. Attardi, F. Simi, Tanganelli, and A. Tommasi, "Learning conceptual descriptions of categories," *Rapporto Tecnico, Dipartimento di Informatica,* TR-99-21, 1999.

[2] P. K. Chan, "A non-invasive learning approach to building web user profiles," *KDD-99 Workshop on Web Usage Analysis and User Profiling*, 1999.

[3] B. Choi and X. Peng, "Dynamic and hierarchical classification of web pages," *Online Information Review*, 28(2) 139-147, 2004.

[4] B. Choi and Z. Yao, "Web page classification," Book Chapter on *Recent Advances in Data Mining and Granular Computing*, Springer-Verag, (in print), 2004.

[5] L. Gravano, H. Garcia-Molina, and A. Tomasic, "Text-source discovery over the internet," *ACM Transactions on Database Systems*, 24(2) 229-264, June 1999.

[6] D. Grossman, D. Frieder, O. D. Holmes, and D. Roberts, "Integrating structured data and text: A relational approach," *Journal of the American Society for Information Science*, 48(2). 1997.

[7] W. Hsu, and S. Lang, "Classification algorithms for NETNEWS articles," *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management*, 1999.

[8] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," *International Conference on Machine Learning* (ICML), 1997.

[9] T. Joachims, "Text categorization with support vector Machines: Learning with many relevant features," *European conference on Machine Learning (ECML)*, 137-142, Berlin, 1998.

[10] A. Kehagias, V. Petridis, V.G. Kaburlasos, and P. Fragkou, "A comparison of word- and sense-based text categorization using several classification algorithms," *Journal of Intelligent Information Systems*, 21(3), 2001.

[11] D. Koller, and M. Sahami, "Hierarchically classifying documents using very few words," *Proceedings of the 14th international Conference on Machine Learning ECML98*, 1998.

[12] Y. Labrou, and T. Finin, "Yahoo! As an ontology – using Yahoo! categories to describe document," *CIKM '99. Proceedings of the Eighth International Conference on Knowledge and Information Management*, 180-187, 1999.

[13] K. Lang, "Newsweeder: learning to filter news," *Proceedings of the 12th International Conference on Machine Learning*, 331-339,1995.

[14] G. A. Miller, "Nouns in WordNet: a lexical inheritance system," *International Journal of Lexicography*, 3 (4) 1990.

[15] G. A. Miller, R. Beckwith, C. Felbaum, D. Gross, and K. Miller, "Introduction to WordNet : an on-line lexical database," *International Journal of Lexicography*, 3, (4) 235 – 244, 1990.

[16] D. Mladenic, *Machine Learning on non-homogeneo, distributed text data*, PhD thesis, University of Ljubljana, Slovenia. 1998

[17] P. Pantelm and D. Lin, "Discovering word senses from text," *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2002.

[18] B. M. Rodriguez, J. M. Gmezhidalgo, and B. Daz Agudo, "Using WordNet to complement training information in text categorization," *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 1997.

[19] S. Scott and S. Matwin, "Text classification using WordNet hypernyms," *Coting-ACL'98 workshop: usage of WordNet in natural language processing systems*, 45-51, August 1998.

[20] A. Sun, E. Lim, and W. Ng, "Web classification using support vector machine," *WIDM'02*, 2002.

[21] WordNet, http://www.cogsci.princeton.edu/~wn/

[22] D. Yarowsky, "Word sense disambiguation using. statistical models of roget's categories trained on large corpora," *Proceedings of COLING.* Nantes, France, 454-460, 1992.