Feature Selection Based on Semantics

Stephanie Chua
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia.
chlstephanie@fit.unimas.my

Narayanan Kulathuramaiyer
Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia.
nara@fit.unimas.my

Abstract – **The need for an automated text categorization system is spurred on by the extensive increase of digital documents. This paper looks into feature selection, one of the main processes in text categorization. The feature selection approach is based on semantics by employing WordNet [1]. The proposed WordNet-based feature selection approach makes use of synonymous nouns and dominant senses in selecting terms that are reflective of a category's content. Experiments are carried out using the top ten most populated categories of the Reuters-21578 dataset. Results have shown that statistical feature selection approaches, Chi-Square and Information Gain, are able to produce better results when used with the WordNet-based feature selection approach. The use of the WordNet-based feature selection approach with statistical weighting results in a set of terms that is more meaningful compared to the terms chosen by the statistical approaches. In addition, there is also an effective dimensionality reduction of the feature space when the WordNet-based feature selection method is used.**

I. Introduction

The task of document categorization is being carried out everyday. In today's computerized environment, many categorization tasks are still being done manually. This is due to the fact that most digitized documents are in the natural language. Therefore, those documents need to be preprocessed beforehand in order to make it understandable by the computers. The task of preprocessing the documents involves many processes, among which, one of the most significant process is the feature selection process. The feature selection process involves selecting a subset of keywords in a category to represent the category in the categorization task. Feature selection based on statistical approaches is commonly used. However, these approaches do not take into consideration the semantics of the natural language. Almost all our everyday documents are in the natural language. Therefore, in order to categorize them more effectively, there is a need to have the semantics component in the feature selection process.

In this research, we explore the hypothesis that incorporating semantics knowledge into feature selection can improve categorization accuracy and identify keywords that best describe a particular category. In the works that are carried out in this research, we attempt to explain how text categorization can be made more effective by incorporating WordNet as the semantics database in the feature selection process.

In Section II, we give an overview of text categorization. Section III will discuss statistical and semantics feature selection. An introduction to WordNet will be given in Section IV. Section V will briefly describe categorical sense disambiguation. Section VI will give an overview of the approach used for feature selection based on WordNet. Section VII will emphasize the dimensionality reduction achieved in this research. In Section VIII, the experiments are described and the results and analysis are presented in Section IX. Finally Section X concludes the paper with a summary.

II. Text Categorization

Text categorization is defined as assigning new documents to a set of pre-defined categories based on the classification patterns suggested by a training set of categorized documents. Automated text categorization is a field that has been around since the early 1960s [2]. In those days, categorization of text was done manually by constructing classifiers using some knowledge engineering techniques. In other words, it was done by gathering the knowledge of domain experts and then defining a set of rules that incorporate the experts' knowledge in categorizing the documents into a given set of categories. No doubt this technique is time consuming especially if the amount of documents is abundant. With a steep increase in digital documents over the years, manual categorization proves to be inefficient.

As the paradigms shifted in this computer age, the machine learning approach to text categorization starts to gain popularity. Many machine-learning schemes have been applied to text categorization and among them are Naïve Bayes [3], support vector machines (SVM) [4], decision trees [5] and so on. The application of machine learning in the field of text categorization only emerged in the 1990s. The concept behind machine learning in the task of categorization is generally described as a learner that automatically builds a classifier by learning from a set of documents that has already been classified by experts [2].

In this research, we apply the machine learning approach to text categorization. Fig. 1 shows the framework of the text categorization process. Our research focus is on the feature selection process to improve the effectiveness of the text categorization process.

III. Feature Selection: Statistical Vs Semantics

Feature selection is performed in text categorization to tackle the problem of the large dimensionality of the feature space. This process involves selecting a subset of features from the

K. Elleithy (ed.), *Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering,* 471–476.

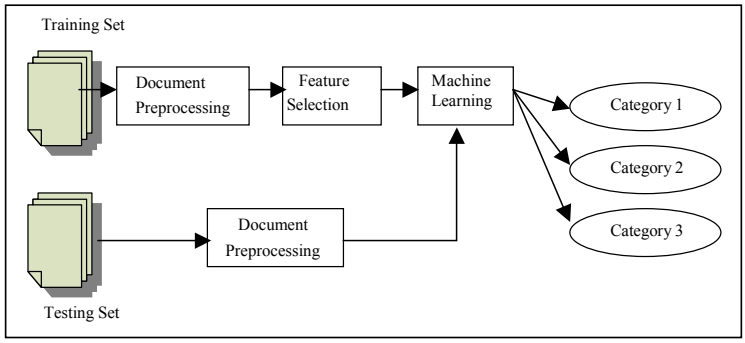

Fig. 1. The text categorization framework

feature space to represent the category. A feature space can contain thousands of features; however, it is not computationally efficient to process a large feature space. A good feature selection approach needs to be employed to select the most suitable features for category representation. There are a number of feature selection approaches, which over the years, are used in a wide range of text categorization tasks. The more widely used ones are statistical-based approaches, which will be discussed in the next section.

A. Statistical Feature Selection

Among the more widely used statistical-based feature selection approaches are Information Gain (IG) [2], [6], [7], [8] and Chi-square (Chi2) [2], [6], [8], [9]. Both IG and Chi2 can reduce the dimension of the vector space by a factor of 100 with no loss of categorization effectiveness [10]. It is thus desirable to develop feature selection approaches with a performance comparable to both IG and Chi2.

1. Information Gain (IG)

Information Gain (IG) or more popularly known as InfoGain, is a feature selection approach that makes use of the presence and absence of a term in a document to select its features. It is frequently used as a term-goodness criterion in the field of machine learning. The number of bits of information is measured for category prediction by using the knowledge of the presence and absence of a term. The amount of information term t_k contains about category c_i is measured and terms that are more indicative of a category based on their presence or absence are chosen.

For each unique term in the training set, information gain is computed and those terms that are above a predetermined threshold are selected as features. A term with a high information gain indicates that it is a good feature for category prediction. The formula of IG is shown in (1).

$$\sum_{c\in\{ci,\bar{ci}\}}\sum_{t\in\{tk,\bar{tk}\}} \mathrm{P}(t,c)\cdot\log 2\frac{\mathrm{P}(t,c)}{P(c).P(t)} \qquad (1)$$

2. Chi-Square (Chi2)

This method measures the degree of dependence between a term and a category. If a term is independent of a category, it will have a value of zero. A low value of Chi2 signifies a high degree of independence of term t_k and category c_i while a high value shows otherwise. A term with a high value of Chi2 shows that it is more dependant on a category and is therefore, more likely to be added to the feature space. These highly dependent features are selected because of their discriminating power. For each category, the Chi2 value for each unique term is computed and those terms that are below a predefined threshold are removed from the feature space. The formula of Chi2 is shown in (2).

$$\frac{\left[\mathrm{P}(t\,k,ci)\cdot\mathrm{P}(\bar{t}k,\bar{ci})-\mathrm{P}(t\,k,\bar{ci})\cdot\mathrm{P}(\bar{t}k,ci)\right]^2}{\mathrm{P}(t\,k)\cdot\mathrm{P}(\bar{t}k)\cdot\mathrm{P}(ci)\cdot\mathrm{P}(\bar{ci})} \qquad (2)$$

B. Semantics Feature Selection

Semantics feature selection is not as widely explored as the statistical approaches. It refers to the selection of features based on its semantics value. Semantics is the study of word meanings. Digital dictionaries and word databases are commonly used to handle the linguistics aspects of text documents. Works by [11] investigated the usage of cascaded feature selection (CFS) in SVM text categorization. Their work highlights the potential of making use of synonyms in feature selection. Their approach shows promising results. Further to that, they also explore the use of parts-of-speech (POS) in a variable CFS [12]. Here, they use a two-step POS selection for SVM based text categorization.

In this research, WordNet, a lexical database, will be used to add in semantics information in the feature selection process. Unlike statistical feature selection, in semantics feature selection, a feature is chosen based on its semantics content, rather than based on its statistical value.

IV. Introduction To WordNet

WordNet is an online thesaurus and an online dictionary. It can be considered as a dictionary based on psycholinguistics principles. WordNet contains nouns, verbs, adjectives and adverbs as parts-of-speech (POS). Function words are omitted based on the notion that they are stored separately as part of the syntactic component of language [1].

WordNet is organized by relations such as synonym, antonym, hyponym/hypernym and holonym/meronym. While synonym and antonym are lexical relations between word forms, both hyponym/hypernym and holonym/meronym are semantics relations between word meanings. Generally, synonyms are words having the same meaning and antonyms are words having opposite meanings. On the other hand, semantics pointers include "IS-A", "PART-OF/HAS-PART", "MEMBER-OF/HAS-MEMBER" and "SUBSTANCE-OF/HAS-SUBSTANCE" [1], [13]. The "IS-A" relationship is also known as the hyponym/hypernym relationship, where hyponym is the subset and hypernym is the superset. "PART-OF/HAS-PART", "MEMBER-OF/HAS-MEMBER" and "SUBSTANCE-OF/HAS-SUBSTANCE" is also known as the holonym/meronym relationship. Holonym is the inverse of meronym where, if x is a holonym of y, then y is a meronym of x.

The information in WordNet is organized into sets of words called synsets. Each synset in WordNet has a unique signature that differentiates it from other synsets. Each of the synset contains a list of synonymous words and semantics pointers that illustrate the relationships between it and other synsets.

In this research, WordNet is chosen over other alternatives, as there are a few advantages of WordNet that can be exploited. First and foremost, it links related words in a structure defined as a synset. Words are ordered hyponymically, that is, they are grouped and sorted in a hierarchy based on their meanings. Different concepts are represented by different synsets [1]. Besides that, it is able to provide semantics information, consistently structured and electronically available.

V. Categorical Sense Disambiguation

WordNet contains a list of senses for each of the words in its dictionary. Therefore, WordNet is able to provide each word with a list of senses that it has. By looking at the context of a word in a category, WordNet can be used to provide the sense of the word. When two synsets overlap, the sense of each of the corresponding term is identified.

Although categorical word senses are identified in this research, the sense information is not used. We merely use the sense information to identify terms with dominant senses by finding the overlapping synset sense signatures. Further processing is required to incorporate the actual word sense of each noun. Therefore, categorical sense disambiguation is applied only to determine the sense of the synonymous terms for a category.

VI. WordNet-Based Feature Selection

In the WordNet-based feature selection approach, only nouns are considered. Preliminary experiments indicate that the use of other parts-of-speech (POS) does not significantly enhance performance. Therefore, the nouns are first identified based on the nouns in the WordNet's dictionary. Synonyms that co-exist in a category are cross-referenced with the help of WordNet's dictionary. The terms obtained from cross-referencing will be the features that will be used to represent a category. This approach is illustrated in Fig. 2.

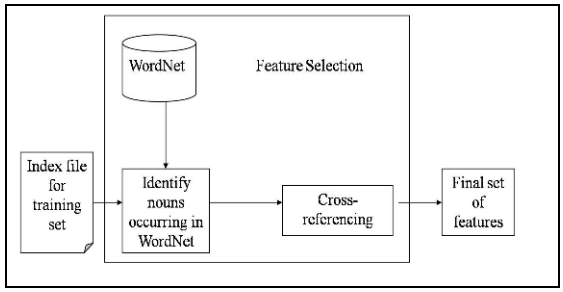

Fig. 2. The WordNet-based feature selection approach

The use of the WordNet-based approach allows us to determine whether semantics feature selection can enhance the quality of features for automated text categorization.

The difference between the semantics approach and the statistical approach is that, in the semantics approach, synonyms are chosen as features and are then weighted using Chi2 and IG. Our research makes use of the Chi2 and IG formulae from the works of [6].

With the WordNet-based approach for feature selection, insignificant words and noise can be filtered. These insignificant words consist of non-English words, wrongly spelt words, insignificant abbreviations and names. Terms like "govodi" and "pik" are actually meaningless in representing a category. Statistical approaches like Chi2 and IG do not take into consideration whether a term is misspelt or is reflective of a category. By using the WordNet-based approach for feature selection, we can actually tackle this problem by filtering these terms and at the same time, make use of the available synonym relationship and word senses in WordNet to identify semantics features in a category. The WordNet-based feature selection approach is able to choose a set of terms that is more reflective of a category's content. This is in line with inducing a classifier that will act more like a human expert rather than having a classifier rely only on statistical findings.

The example below illustrates the approach used. We will look at all the senses for four nouns; "corn", "maize", "acquisition" and "ship". Each sense has a signature, which is referred to as a synset. Every synset contains synonyms to reflect a sense.

TABLE I
LIST OF TERMS AND THEIR SYNSETS FOR EACH SENSE

Terms	Synsets for all senses
Corn	**Sense 1: {corn, maize, Indian corn, Zea mays}** Sense 2: {corn} Sense 3: {corn, edible corn} Sense 4: {corn, clavus} Sense 5: {wheat, corn} Sense 6: {corn whiskey, corn whisky, corn}
Maize	**Sense 1: {corn, maize, Indian corn, Zea mays}** Sense 2: {gamboge, lemon, lemon yellow, maize}
Acquisition	Sense 1: {acquisition} Sense 2: {acquisition} Sense 3: {learning, acquisition} Sense 4: {skill, accomplishment, acquirement, acquisition, attainment}
Ship	Sense 1: {ship}

From Table I, two identical synsets are identified (indicated in bold). They are sense 1 of "corn" and sense 1 of "maize", which have the same synset signatures. Thus, the terms "corn" and "maize" will be selected as terms to represent a category in feature selection. The use of categorical sense disambiguation is employed here to automatically disambiguate semantically related terms. The dominant senses for terms in each category can be determined by cross-referencing to find identical synsets with the same signatures.

VII. DIMENSIONALITY REDUCTION

The WordNet-based feature selection approach is also effective in reducing the dimensionality of the feature space. Table II shows the percentage of terms reduction when the approach is used compared to the number of unique terms in each category. Generally, the WordNet-based approach for feature selection is able to reduce the number of terms by more than 67%.

TABLE II
PERCENTAGE OF TERMS REDUCTION FOR THE REUTERS-21578 TOP TEN CATEGORIES

Category	No. of unique terms in each category	No. of terms selected by the WordNet-based approach	Percentage of terms reduction (%)
Acq	10760	2793	74.0
Corn	3089	955	69.1
Crude	5890	1834	68.9
Earn	10152	2342	76.9
Grain	5231	1716	67.2
Interest	3679	1147	68.8
Money-fx	5323	1667	68.7
Ship	3902	1258	67.8
Trade	5569	1823	67.3
Wheat	3358	1109	67.0

The effective reduction displays the ability of the WordNet-based approach to reduce noise while preserving the original contextual information of the documents.

VIII. EXPERIMENTS

The experiments that are carried out are to test and compare the effectiveness of semantics feature selection and statistical feature selection. The dataset used is the Reuters-21578 top ten most populated categories. The experiments are carried out using the Waikato Environment for Knowledge Analysis (WEKA) [14] machine learning tool, applying the multinomial Naïve Bayes machine learning scheme.

Experiments are carried out to compare and contrast the following feature selection approaches; Information Gain (IG), Chi-square (Chi2), WordNet-based feature selection using IG weighting (W-IG) and WordNet-based feature selection using Chi2 weighting (W-Chi2). The formulae for IG and Chi2 are obtained from [6].

Chi2 and IG are chosen in this research to be used for comparison because previous experiments by other researchers have proven that these approaches are successfully implemented as statistical feature selection approaches. Therefore, by making use of these two approaches as benchmark for comparison, we will be able to see how well the proposed WordNet-based feature selection approach can perform.

There are two sets of experiments. Each set of experiments consists of 6 different term sizes: 10, 20, 50, 100, 200 and 500. These different term sizes were chosen to evaluate the effectiveness of the classifier to see which term size can give the optimal performance for the classifier. The two sets of experiments are:

1. Comparison between Chi2 and WordNet-based feature selection using Chi2 weighting (W-Chi2).
2. Comparison between IG and WordNet-based feature selection using IG weighting (W-IG).

The aim of these experiments is to determine the effectiveness of the WordNet-based feature selection approach.

IX. RESULTS AND ANALYSIS

The F_1 measure with micro-averaged scores across all the ten categories is used in measuring the results. F_1 measure combines both the value of precision (P) and recall (R) to give a more effective result to indicate the effectiveness of the classifier's performance. The formula of F_1 measure is given in (3).

$$F_1 = 2 \cdot P \cdot R / (P + R) \quad (3)$$

Fig. 3 and 4 shows the micro-averaged F_1 measure of Chi2 and W-Chi2, as well as, IG and W-IG, for the Reuters-21578 top ten categories.

From Fig. 3 and 4, it is noted that both W-Chi2 and W-IG is able to perform better than the statistical approach itself, with the exception at term size 10. At all other term size thresholds,

there is a slight increase in the F_1 measure value when the WordNet-based approach is used with the statistical weighting. The reason for this is that the WordNet-based approach needs a larger number of features to capture adequate representative categorical features, as well as, semantics information. This is the reason why it did not improve on the results of the statistical approaches at term size 10, while across all other term sizes, the WordNet-based approach with the statistical weighting is able to produce some improvements over the statistical approach itself.

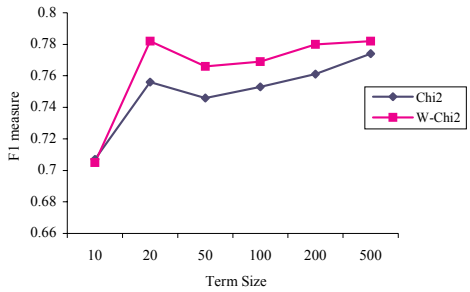

Fig 3. Comparison of micro-averaged F_1 measure between Chi2 and W-Chi2 for the Reuters-21578 top ten categories

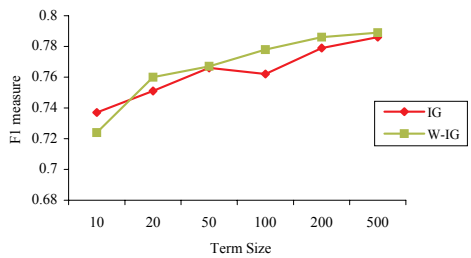

Fig. 4. Comparison of micro-averaged F_1 measure between IG and W-IG for the Reuters-21578 top ten categories

By using the WordNet-based approach for feature selection, a set of terms that is more meaningful and more reflective of a category's content can be obtained. We illustrate this by using two examples. The first example is the top 20 terms chosen by Chi2 and W-Chi2 for the category "acquisition" (acq), which is listed in Table III.

TABLE III
THE TOP 20 TERMS CHOSEN BY CHI2 AND W-CHI2 FOR THE CATEGORY "ACQ"

Feature selection approach	Top 20 terms for category "acq"
Chi2	Shares, Offer, Lt, Stake, Merger, Cts, Acquisition, Company, Inc, Acquire, Net, Loss, Corp, Usair, Common, Mln, Unit, Shr, Stock, Sell
W-Chi2	Shares, Offer, Stake, Merger, Cts, Acquisition, Company, Net, Loss, Corp, Common, Unit, Stock, Sell, Buy, Takeover, Shareholders, Trade, Transaction, Bid

When the list of terms chosen by Chi2 and W-Chi2 is compared to each other, it is noted that there are six terms that differ. These six terms differentiate the results between the two approaches. The six terms for each approach are listed in Table IV.

TABLE IV
THE SIX TERMS THAT DIFFERENTIATE BETWEEN CHI2 AND W-CHI2

Chi2	W-Chi2
Lt	Buy
Inc	Takeover
Acquire	Shareholders
Usair	Trade
Mln	Transaction
Shr	Bid

If human experts were asked to choose a set of terms to represent the category "acq", it is very likely that they would choose the terms listed under W-Chi2 in Table IV. All the terms under W-Chi2 clearly reflect the concept of acquisition. Under Chi2, there is only one term that strongly represents the concept of acquisition, which is "acquire". With W-Chi2, the term "acquire" is not chosen simply because the approach only consider nouns and not verbs.

The second example is the top 20 terms chosen by IG and W-IG for the category "wheat", which is listed in Table V.

TABLE V
THE TOP 20 TERMS CHOSEN BY IG AND W-IG FOR THE CATEGORY "WHEAT"

Feature selection approach	Top 20 terms for category "wheat"
IG	Wheat, Tonnes, Vs, Lt, Agriculture, Net, Export, Loss, Soviet, Grain, Crop, Winter, Usda, Department, Company, Bank, Barley, Lyng, Eep, Program
W-IG	Wheat, Tonnes, Agriculture, Net, Export, Loss, Grain, Crop, Department, Company, Program, Farm, Profit, Subsidy, Share, Farmers, Shares, Tonne, Commodity, Corn

From the comparison of the list of terms chosen by IG and W-IG, it is noted that there are nine terms that differ. These are the nine terms that differentiate the results between the two approaches. The nine terms for each approach are listed in Table VI.

TABLE VI
THE NINE TERMS THAT DIFFERENTIATE BETWEEN IG AND W-IG

IG	W-IG
Vs	Farm
Lt	Profit
Soviet	Subsidy
Winter	Share
Usda	Farmers
Bank	Shares
Barley	Tonne
Lyng	Commodity
Eep	Corn

Again, it is seen in Table VI that W-IG has terms that closely represent the category "wheat" as compared to the terms chosen by IG. Both Chi2 and IG will include terms that are statistically significant regardless of whether they are reflective of the category or not. For example, in Table IV and VI, we can see that both Chi2 and IG choose the word "Lt". As a human being would think, this word bears no connection to both categories "acq" and "wheat". It is not meaningful to both the categories. Therefore, with the use of the WordNet-based approach for feature selection, it is seen that a set of terms that is more meaningful and more reflective of a category's content can be obtained.

X. Conclusion

In this research, it has been shown that there is another approach for feature selection other than the statistical approach. The semantics feature selection is seen as a promising approach for feature selection, as it is able to select features that are more meaningful and more reflective of a category's content. It is also observed that when this approach is used with the statistical weighting, it can perform better than the statistical approach itself with improvements in categorization accuracy. Apart from that, this approach is also effective in reducing the dimensionality of the feature space. To summarize, this research has demonstrated the ability to extract meaningful terms from statistical features. It could thus be applied as a means to filter terms and potentially lead towards better text understanding. In conclusion, the incorporation of the semantics component using WordNet is capable of improving the effectiveness of tasks that involves the natural language.

References

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

[2] Sebastiani, F. (1999). A tutorial on automated text categorization. In *Proceedings of ASAI-99, 1st Argentinean Symposium on Artificial Intelligence* (Analia Amandi and Ricardo Zunino, eds), pp 7 – 35, Buenos Aires, AR.

[3] McCallum A. and Nigam, K. (1998). A comparison of event model for naive Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*.

[4] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *In Proceedings of the Tenth European Conference on Machine Learning (ECML)*, pp 137 – 142.

[5] Holmes G. and Trigg L. (1999). A diagnostic tool for tree based supervised classification learning algorithms. In *Proceedings of the Sixth International Conference on Neural Information Processing (ICONIP)*, Perth, Western Australia, Volume II, pp 514 – 519.

[6] Debole, F. and Sebastiani, F. (2003). Supervised term weighting for automated text categorization. In *Proceedings of SAC-03, 18th ACM Symposium on Applied Computing*, Melbourne, AS, pp 784 – 788.

[7] Mladenić, D. (1998). Turning Yahoo into an automatic web-page classifier. In *Proceedings of the 13th European Conference on Aritficial Intelligence (ECAI)*, pp 473 – 474.

[8] Yang Y. and Pedersen, J. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML 97)*, pp 412 – 420, Nashville, TE, USA.

[9] Schütze, H., Hull, D. A. and Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in rieval*, Se*Information Ret*attle, US, pp 229 – 237.

[10] Yang, Y. (1999). *An Evaluation of Ststistical Approaches to Text Categorization*. Information Retrieval, 1-2(1):69-90.

[11] Masuyama, T. and Nakagawa, H., "Applying Cascaded Feature Selection to SVM Text Categorization", in the DEXA Workshops, 2002, pp. 241-245.

[12] Masuyama, T. & Nakagawa, H., Two step POS selection for SVM based text categorization, in *IEICE Transaction on Information System*, Vol. E87-D, No.2, February 2004.

[13] Richardson, R., Smeaton, A. and Murphy, J. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. In *Proceedings of AICS Conference.* Trinity College, Dublin.

[14] Witten, I. H. and Frank, E. *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.