USING WORDNET TO ENHANCE FEATURE SELECTION
IN AUTOMATED TEXT CATEGORIZATION

Stephanie Chua Hui Li

Master of Science
(Information Technology)
2004
USING WORDNET TO ENHANCE FEATURE SELECTION IN AUTOMATED TEXT CATEGORIZATION

P. KHIDMAT MAKLUMAT AKADEMIK

STEPHANIE CHUA HUI LI

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2004
Acknowledgment

First and foremost, I would like to extend my utmost gratitude to my supervisor, Associate Professor Narayanan Kulathuramaiyer for all his invaluable guidance and encouragement throughout this research. I would also like to thank Mr Bong Chih How for his resources, advice and ideas pertaining to this research. My acknowledgment also goes to the Pelan Tindakan Pembangunan Teknologi Perindustrian (PTPTP) for the financial support provided. I also have all my friends to thank for their kindness and understanding as well as the discussions and support given, especially during the difficult times. Finally, I am deeply grateful to my beloved family for their unconditional love, patience, support and encouragement.
Abstract

In the field of automated text categorization, the large dimensionality of the feature space is a major problem as it involves extensive computations. Feature selection is one of the approaches to reduce the dimensionality of the feature space. This research explores the use of WordNet (Miller et al., 1990), a lexical database, for performing feature selection for an automated text categorization system. The WordNet-based approach employs lexical and semantics information for feature selection. WordNet allows the selection of terms that are lexically and semantically representative of a category of documents, as opposed to statistical approaches traditionally used for feature selection.

We proposed three WordNet based approaches for feature selection. The first one is to use the WordNet nouns approach that selects all nouns in WordNet that occur in each category as features. The second approach is based on lexical semantics that selects synonymous terms that co-occur in a category while the third approach is a combination of the lexical semantics approach with statistical feature selection methods.

The lexical semantics approach performed better than the WordNet nouns approach with more than 40% of reduction in feature space in the experiments using the Reuters-21578 dataset. The lexical semantics approach also outperformed popular statistical feature selection methods, namely, Chi-Square (Chi2) and Information Gain (IG). The combined approach has improved the performance of the statistical methods. WordNet has successfully been used to enhance feature selection,
highlighting the possibility of determining semantic features automatically. The limitations of the lexical semantics approach are also highlighted, proposing an improved framework and an extension to overcome them.
Abstrak


Kaedah leksikal semantik adalah lebih efektif berbanding dengan kaedah kata nama WordNet dengan lebih daripada 40% pengurangan saiz dimensi perkataan dalam eksperimen yang menggunakan set data Reuters-21578. Kaedah leksikal semantik juga
berjaya memperolehi keputusan yang lebih baik daripada kaedah statistikal popular, iaitu, Chi-square (Chi2) dan Information Gain (IG). Kaedah kombinasi telah berjaya mempertingkatkan keputusan yang dicapai oleh kaedah statistikal. WordNet telah berjaya digunakan untuk mempertingkatkan kualiti proses pemilihan teks dengan kaedah leksikal semantik ini lebih efektif berbanding dengan kaedah statistikal yang sedia ada, iaitu, Chi2 dan IG. Dengan itu, kaedah ini dilihat sebagai satu kaedah pemilihan teks yang berpotensi. Limitasi kaedah leksikal semantik ini telah dikenalpasti dan rangka kerja yang lebih baik serta lanjutan untuk mengatasi limitasi kaedah tersebut telah dicadangkan.
# Table of Contents

<table>
<thead>
<tr>
<th>Acknowledgment</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Abstrak</td>
<td>iv</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>vi</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xiv</td>
</tr>
</tbody>
</table>

## Chapter 1  Introduction

1.1 Introduction to Text Categorization 1

1.2 Machine Learning Approach to Text Categorization 2

1.3 Framework for Automated Text Categorization 3

1.4 Dimensionality Problem 5

1.5 Feature Selection vs Feature Extraction 6

1.5.1 Statistical Feature Selection 7

1.5.2 Natural Language Processing (NLP) Approach to Feature Selection 8

1.6 Problem Statement 9

1.7 Research Objectives 10

1.8 Scope 11

1.9 Chapters Overview 12
Chapter 2 Literature Review

2.1 Introduction

2.2 Text Representation for Text Categorization

2.3 Automatic Approach to Text Categorization
   2.3.1 Knowledge Engineering
   2.3.2 Machine Learning

2.4 Machine Learning Approach to Text Categorization
   2.4.1 Naive Bayes (NB) Classifier
   2.4.2 Single-label vs Multi-label Categorization

2.5 Dimensionality Reduction by Feature Extraction
   2.5.1 Term Clustering
   2.5.2 Latent Semantic Indexing (LSI)

2.6 Dimensionality Reduction by Feature Selection
   2.6.1 Information Gain (IG)
   2.6.2 Chi-Square (Chit)
   2.6.3 Discussion on Information Gain and Chi-Square

2.7 Word Sense Disambiguation

2.8 Introduction to WordNet

2.9 Incorporating WordNet in the Task of Text Categorization
   2.9.1 WordNet in Indexing
   2.9.2 WordNet in Text Representation
   2.9.3 WordNet in Semantic Distance Measurement
   2.9.4 WordNet in Feature Selection

2.10 Comparison and Discussion

2.11 Evaluation Measures
Chapter 3 Design

3.1 Introduction 48
3.2 Text Categorization Framework 49
  3.2.1 Document Preprocessing 50
  3.2.2 Feature Selection 52
    3.2.2.1 WordNet Nouns Approach 52
    3.2.2.2 WordNet Lexical Semantics Approach 53
    3.2.2.3 Categorical Sense Determination 55
    3.2.4 Combined Approach 56
  3.2.3 Machine Learning for Automated Text Categorization 57
3.3 Summary 60

Chapter 4 Implementation

4.1 Introduction 61
4.2 Implementation 61
  4.2.1 Document Preprocessing 64
  4.2.2 Feature Selection 65
    4.2.2.1 WordNet Nouns Approach 65
    4.2.2.2 WordNet Lexical Semantics Approach 66
    4.2.3 Combined Approach 67
  4.2.3 Machine Learning 68
Chapter 5 Results and Analysis

5.1 Introduction 73

5.2 Dataset 73

5.3 Experiments 74

5.4 Results and Analysis for Reuters-21578 Dataset 76

5.4.1 Dimensionality Reduction 76

5.4.2 Experiment 1: WN vs WLS 77

5.4.2.1 F-measure Analysis 77

5.4.2.2 Dimensionality Reduction 79

5.4.2.3 Qualitative Analysis 80

5.4.3 Experiment 2: Comparison between WN, WLS, Chi2 and IG 81

5.4.3.1 F-measure Analysis 81

5.4.3.2 Qualitative Analysis 86

5.4.3.3 Discussion 89

5.4.4 Experiment 3: Comparison between Chi2, IG and Combined Approaches 92

5.4.4.1 F-measure Analysis 92

5.4.4.2 Discussion 94

5.5 Summary 95
Chapter 6 Summary and Future Works

6.1 Introduction 97
6.2 Research Contributions 97
6.3 Limitations 99
6.4 Future Works 100
  6.4.1 Concepts Identification Based on WordNet's Hypernyms 100
  6.4.2 Synonymization Based on WordNet's Synsets 104
  6.4.3 Improved Framework for WordNet based Automated Text Categorization 105
6.5 Conclusion 106

List of References 107

Appendices

APPENDIX A – Comparison of WLS (with singular and plural nouns) vs WLS (with singular nouns only) 117
APPENDIX B – Stop Words List 118
APPENDIX C – Index File Format 124
APPENDIX D – Comparison of WLS (nouns only) vs WLS (all POS) 127
APPENDIX E – ARFF File Format 128
APPENDIX F – Correlation Test Results 132
APPENDIX G – Extract of Codes 134
APPENDIX H – Conference Paper Accepted for the 2nd International Conference on Artificial Intelligence in Engineering and Technology (ICAIET 2004), Kota Kinabalu, Sabah, 3rd – 5th August, 2004 136
APPENDIX I – Conference Paper Accepted for the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), Beijing, China, 20th – 24th September, 2004
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1.1:</td>
<td>Framework for automated text categorization</td>
<td>5</td>
</tr>
<tr>
<td>Figure 2.1:</td>
<td>A HCG showing the concept &quot;car&quot; (extracted from Richardson and Smeaton, 1995)</td>
<td>38</td>
</tr>
<tr>
<td>Figure 3.1:</td>
<td>Text categorization framework</td>
<td>50</td>
</tr>
<tr>
<td>Figure 3.2:</td>
<td>Document preprocessing</td>
<td>51</td>
</tr>
<tr>
<td>Figure 3.3:</td>
<td>An extract of the Reuters-21578 index file</td>
<td>51</td>
</tr>
<tr>
<td>Figure 3.4:</td>
<td>The WordNet nouns approach</td>
<td>53</td>
</tr>
<tr>
<td>Figure 3.5:</td>
<td>The WordNet lexical semantics approach</td>
<td>54</td>
</tr>
<tr>
<td>Figure 3.6:</td>
<td>The combined approach</td>
<td>57</td>
</tr>
<tr>
<td>Figure 3.7:</td>
<td>The machine learning approach for text categorization</td>
<td>59</td>
</tr>
<tr>
<td>Figure 4.1:</td>
<td>Context diagram for the text categorization process</td>
<td>63</td>
</tr>
<tr>
<td>Figure 4.2:</td>
<td>Child diagram for process 2 – Create index file for dataset</td>
<td>65</td>
</tr>
<tr>
<td>Figure 4.3:</td>
<td>Child diagram of process 6 – Cross-referencing</td>
<td>67</td>
</tr>
<tr>
<td>Figure 4.4:</td>
<td>Examples of synsets</td>
<td>67</td>
</tr>
<tr>
<td>Figure 4.5:</td>
<td>Child diagram of process 4 – Perform statistical feature selection</td>
<td>68</td>
</tr>
<tr>
<td>Figure 4.6:</td>
<td>Child diagram for process 9 – Create ARFF files</td>
<td>70</td>
</tr>
<tr>
<td>Figure 4.7:</td>
<td>An extract of an ARFF file</td>
<td>70</td>
</tr>
<tr>
<td>Figure 4.8:</td>
<td>Screenshot of the output from the machine learning process</td>
<td>72</td>
</tr>
<tr>
<td>Figure 5.1:</td>
<td>Macro-averaged f-measure comparison between WLS and WN for the Reuters-21578 dataset</td>
<td>78</td>
</tr>
</tbody>
</table>
Figure 5.2: Micro-averaged f-measure comparison between LS and WN for the Reuters-21578 dataset

Figure 5.3: Macro-averaged f-measure for the Reuters-21578 dataset

Figure 5.4: Micro-averaged f-measure for the Reuters-21578 dataset

Figure 5.5: Macro-averaged f-measure for the Reuters-21578 top ten categories

Figure 5.6: Micro-averaged f-measure for the Reuters-21578 top ten categories

Figure 5.7: Comparison of the performance of WLS, Chi2 and IG in different categories at term size 100 for the Reuters-21578 top ten categories

Figure 5.8: Comparison of micro-averaged f-measure between Chi2 and WLS-Chi2 for the Reuters-21578 top ten categories

Figure 5.9: Comparison of micro-averaged f-measure between IG and WLS-IG for the Reuters-21578 top ten categories

Figure 6.1: List of 25 unique beginners for WordNet nouns (extracted from Miller et al., 1990)

Figure 6.2: Diagrammatic representation of hyponymic relations among seven unique beginners denoting different kinds of tangible things (extracted from Miller et al., 1990)

Figure 6.3: Proposed framework to improve automated text categorization using the WordNet-based approaches
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Mathematical formula for feature selection methods (extracted from Debole and Sebastiani, 2003)</td>
<td>26</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Criteria for performance of feature selection approaches (extracted from Yang and Pedersen, 1997)</td>
<td>29</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>The semantic relationship in WordNet</td>
<td>32</td>
</tr>
<tr>
<td>Table 2.4</td>
<td>Comparison between works using WordNet</td>
<td>40</td>
</tr>
<tr>
<td>Table 2.5</td>
<td>Contingency table for binary categorization (extracted from Witten and Frank, 2000)</td>
<td>42</td>
</tr>
<tr>
<td>Table 2.6</td>
<td>Computation for the performance metrics (extracted from Sebastiani, 1999)</td>
<td>43</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>List of terms and their synsets in all senses</td>
<td>55</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>The number of categories in each category set and the number of categories with at least 1 occurrence and 20 occurrences</td>
<td>74</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Percentage of terms reduction when the WLS approach is used for the Reuters-21578 top ten categories compared to the number of unique terms in each category</td>
<td>77</td>
</tr>
<tr>
<td>Table 5.3</td>
<td>Percentage of terms reduction when the WLS approach is used for the Reuters-21578 top ten categories compared to the WN approach</td>
<td>79</td>
</tr>
<tr>
<td>Table 5.4</td>
<td>Top 20 terms for category “Interest” for WLS and WN</td>
<td>80</td>
</tr>
<tr>
<td>Table 5.5</td>
<td>Macro-averaged f-measure for the Reuters-21578 dataset</td>
<td>82</td>
</tr>
<tr>
<td>Table 5.6</td>
<td>Micro-averaged f-measure for the Reuters-21578 dataset</td>
<td>82</td>
</tr>
</tbody>
</table>
Table 5.7: Macro-averaged f-measure for the Reuters-21578 top ten categories

Table 5.8: Micro-averaged f-measure for the Reuters-21578 top ten categories

Table 5.9: F-measure for each category for Chi2, IG and WLS for the Reuters-21578 top ten categories

Table 5.10: Top 20 terms for category “Acq” for Chi2, IG and WLS

Table 5.11: Top 20 terms for category “Earn” for Chi2, IG and WLS

Table 5.12: Top 20 terms for category “Interest” for Chi2, IG and WLS

Table 5.13: Comparison of micro-averaged f-measure between Chi2 and WLS-Chi2 as well as IG and WLS-IG for the Reuters-21578 top ten categories

Table 5.14: F-measure for each category for Chi2 and WLS-Chi2 as well as IG and WLS-IG at term size 10 for the Reuters-21578 top ten categories
1.1 Introduction to 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 (Yang and Liu, 1999). Text categorization has been actively pursued since the early 1960s. Classifiers are used to classify documents in text categorization. Previously, classifiers were manually built by knowledge engineering effort. This knowledge engineering approach requires experts to manually build classifiers by defining rules for categorization (refer to Section 2.3.1). Even today, categorization task is still largely carried out manually by human experts. With the increase in digital documents over the years, manual categorization by the experts becomes overwhelming. Therefore, an automated approach to text categorization is very much needed to assist in the categorization tasks.

With the escalating growth of new applications, more researches are now being carried out in the field of automated text categorization. Automated text categorization has been applied in many areas such as document filtering, automatic document indexing and search engines.

There are three approaches to automated text categorization. They are the machine learning approach, the similarity measurement approach and the knowledge-based approach. The machine learning approach uses learning schemes such as rule inducers, decision trees, support vector machines (SVM) and so on to perform
categorization. The similarity measurement approach uses a similarity measurement algorithm to determine the degree of similarity of a new document with the documents in a category to do categorization. The knowledge-based approach obtains knowledge from human expert to categorize documents. In this research, the focus is on the machine learning approach. According to Sebastiani (1999), this approach is the fastest and most effective approach to text categorization.

1.2 Machine Learning Approach to Text Categorization

Machine learning is the study of computer algorithms that automatically progress towards improvement of performance by learning experience. Applications of machine learning include data mining programs that discover general rules in large datasets, information filtering systems that automatically learn users' interests and categorization systems that classify documents into categories.

Machine learning has been applied to the field of text categorization since the early 1990s. In this approach, the classifier for a category is automatically built by learning from the documents in the training set, which has been classified manually by experts. The learner will generate a classifier by studying the characteristics of the documents from the training set and then the classifier will be used to classify new and unseen documents from the testing set. This type of learning is called supervised learning because the learning process is coupled with the knowledge of the categories to which a document belongs in the training set. There is also an unsupervised learning method where the process of learning is performed without any prior category knowledge or information. It is called unsupervised because the learning
process is not guided by category labels attached to the documents. This research adopts the supervised machine learning framework to exploit the available manually-classified documents.

1.3 Framework for Automated Text Categorization

The framework for automated text categorization is shown in Figure 1.1. The dataset to be used for text categorization is divided into the training set and the testing set. The training set is used for constructing a classifier through machine learning while the testing set is used for evaluating the effectiveness of the trained classifier. Indexing is the process of cataloging the documents in the training and testing set into a suitable representation of the documents' content. Before indexing is done, document preprocessing is first performed on the dataset. The general processes involved in document preprocessing are stop words removal and stemming. Stop words removal is performed to remove non-significant terms, such as articles and prepositions while stemming is a process of normalizing terms by removing prefixes and suffixes to derive root terms. There are two kinds of stemming; one is based on inflectional morphology while the other is based on derivational morphology. Stemming based on inflectional morphology deals with the change of a word form, usually by removing a suffix for plural words while stemming based on derivational morphology concerns with removing prefixes and suffixes of words to obtain the root form.

Stemming based on derivational morphology was not employed in this research, as this results in stemmed words losing their original meaning. For example, the word
"international" will give the stemmed form of "intern". As such, we have explored the use of stemming based on inflectional morphology, whereby we make use of the WordNet's dictionary to derive root terms from plural forms. The use of stemming in text categorization has been questioned by Sebastiani (1999) as there are studies (Baker and McCallum, 1998) reporting that it decreases the effectiveness in categorization. Moreover, stemming on large and full text database is very time-consuming as pointed out by Korfhage (1997). According to Korfhage, stemming only improves document representation by 5%, and thus, has a lesser impact on full text documents. The result of inflectional stemming as used in our research is shown in Appendix A. Although the number of terms is reduced when plural forms of words are excluded in the feature space, however, the results of using these features is slightly worse than that having the plural forms. This is because no stemming was performed to the test documents prior to the categorization task. Even with a minor reduction in performance, the reduction in feature space is considered significant.

Dimensionality reduction is performed to reduce the number of features in the feature space. Two popular approaches for dimensionality reduction are feature selection and feature extraction. These approaches will be discussed in Section 1.5.

After dimensionality reduction, the selected features will be used to train a classifier. The classifier will learn to categorize documents based on the features provided. After the learning process, the learned classifier could then be used to classify new and unseen documents. The testing set represents new and unseen documents, as this dataset is not involved in the induction of the classifier. In this way, the performance of the classifier is evaluated.
1.4 Dimensionality Problem

The large dimensionality of the feature space is a well-known problem in text categorization. One document can easily contain hundreds to thousands of words and this number increases as the number of documents increases. This creates a problem for the learning process by the induced classifiers, as these classifiers are unable to handle the enormously high number of terms derived from the documents. Therefore, dimensionality reduction has to be performed where the number of terms, $t$, is reduced to $t'$, where $t' < t$.

One advantage of dimensionality reduction is that it reduces the problem of overfitting. Overfitting describes the characteristics of a classifier that is particularly tuned towards the characteristics of the training data. In other words, the classifier is extremely good in classifying the data that it is trained on but is not as good at classifying other data. Overfitting can be avoided when training captures constitutive
features rather than contingent ones. Constitutive features are the terms that are generally representative of a category. On the other hand, contingent features are based on specific patterns that occur in the training set that need not actually be representative of the category and this usually causes the problem of overfitting. Sebastiani (1991) gave an example that, if a classifier for category “Cars for sale” were trained on just three positive examples where two of the examples contained the sale of a yellow car, then the classifier will consider the contingent feature “yellow” as a feature for the category. To avoid overfitting, experiments by Fuhr and Buckley (1991) suggested a ratio of 50 to 100 training examples per feature. However, if dimensionality reduction is performed, overfitting may be avoided even with a smaller set of training examples.

1.5 Feature Selection vs Feature Extraction

Both feature selection and feature extraction are approaches for dimensionality reduction. Feature selection is a process of selecting a subset of features from the feature space. It is performed to select meaningful features, representing the thousands of words in the training set. These selected features, when used in categorization task, should yield the smallest reduction in effectiveness when compared to using the whole feature set.

Feature extraction, on the other hand, is the creation of a new set of terms from the original set. This means that the new set of terms is not contained in the documents but is created from the existing terms to represent a category. Two popular feature extraction approaches are term clustering and latent semantic indexing (LSI). LSI is
a method that compresses the original features in the vector space into a new lower dimensional vector space that consist of the combinations of the original features based on their patterns of co-occurrence in the original vector space. Term clustering tends to deal with the problem of terms redundancy by grouping strongly related words into clusters. LSI has been promising but it is difficult to see why it works well as the features are latent in the semantic structure of the vocabulary used in the corpus (Sebastiani, 1999). A discussion on term clustering and LSI is presented in Section 2.5. In this research, we adopted feature selection to reduce the dimensionality of the vector space. We chose feature selection, as the features derived are made explicit for further analysis.

1.5.1 Statistical Feature Selection

Mathematical formulas are derived from the statistical information of documents and the terms and frequencies in a dataset are used to calculate a weight for each term in a document. The mathematical formulas, ranging from the simple Document Frequency to the more complex Chi-Square (Chi2) and Information Gain (IG) formulae, calculate the weight of each term to determine whether the term should be added to the feature space. Each term that surpasses a threshold value will be added as a feature in the feature space to represent a category. Both Chi2 and IG are feature selection approaches that are purely based on statistics of terms in the dataset. Therefore, these two approaches do not consider the semantics of each term to determine the relevance of the category it represents. Chi2 and IG are further discussed in Section 2.6.