**Kwan Yi and Jamshid Beheshti**
**Graduate School of Library & Information Studies**
**McGill University**

# A Comparative Study on Feature Selection of Text Categorization for Hidden Markov Models

## Abstract

### Background & Purpose

Text categorization (TC) is the task of automatically assigning pre-set categories to textual digital documents. A number of statistical TC models with machine learning techniques have been proposed and reported to tackle the TC problem. One major difficulty with TC models is a high-dimensional feature space. A feature, normally a word or a phrase, is an informative attribute conveying the subject content of a document. Text categorization tasks with a large number of training documents can easily lead to a few thousand features. Since a large number of features, referred to as high dimensionality[*] of feature space, may significantly increase computational times for the TC models, it is highly desirable to reduce the feature space without the loss of model performance.

This paper is a comparative study of feature selection methods in a statistical learning model of text categorization. The dual purposes of this study are to investigate popular dimensionality reduction methods in general TC domain and to probe task-dependent dimensionality reduction methods, for the Hidden Markov Models[1] (HMMs) of the text categorization problem. Our aim is to explore how features should be selected for this model, based on the statistical properties of the related TC task. This study seeks empirical evidence for the following questions: (1) how much do the selected dimensionality reduction methods improve the classification accuracy of the HMM for text categorization? (2) How can the statistical properties of the task be used for reduction methods?

### Conceptual framework

Since the early 1990's, research on TC has been shifted towards the development of statistical learning models using machine learning techniques, such as decision trees[2], Bayesian models[3], and Support Vector Machines[4]. A large number of the dimensionality of feature space for a TC task can not be accommodated in learning models due to the formidable computational times. The reduction of high dimensionality can be done in two different ways. First, a feature-selection approach may reduce the size of the features to be considered by selecting a subset of all available features. Second, a feature-extraction approach may use synthetic features, which do not occur in original documents. Our study focuses on the first approach. Various dimensionality reduction methods have been proposed and tested on different learning models in the TC domain, including Document frequency[5] [6], Information gain[7] [8], Mutual information[9] [10], Chi-square[11] [12], Odds ratio[13], and Relevancy score[14]. Comparison studies of reduction methods have been carried out, and the ranking of reduction techniques on the effectiveness of the method

[*] It refers to the number of distinct features

has been reported[15] [16]. However, as Sebastiani (2002) points out, more comparative studies on diverse experimental settings, such as different classifiers, and different tasks, need to be conducted.

## Methodology

Several popular reduction methods - Information gain, Mutual information, and Chi-Square - and other task-related methods are tested in this study. The research is conducted in three phases: (1) Data collection, training data,[17] and test data[18] (2) Classifier creation, and (3) Classifier Testing.

(1) A set of cataloguing records from the OCLC[19] WorldCat database is collected to constitute a training set for the statistical model. A subset of information in cataloguing records containing topical subjects and their descriptors is used as the training data set in the proposed model. A database for test data sets is created containing digital documents previously classified by professional librarians. In selecting these documents, the availability of content and the type of content is considered. The dissertation abstracts from Proquest Digital Dissertations database (PQDD)[20] has been selected for test set of this system. Library of Congress Classification (LCC) and Library of Congress Subject Headings (LCSH) for the selected abstracts are found in OCLC FirstSearch Database-WorldCat database.

(2) The statistical learning HMMs using different reduction methods are designed and trained (using the sample data from OCLC) as a text classifier. The basic model of this comparison study was designed and built earlier.

(3) Experimental results from different methods are compared to that of the basic classifier. The performance of each classifier will be measured in classification accuracy by comparing the result to the manual classifications by professionals.

* This paper is directly relevant to the theme of *technologies*, as it discusses the improvement of the performance of automatic machine classification system.

**Notes**

[1] Yi, Kwan & Beheshti, Jamshid. 2003. A Text Categorization Model Based on Hidden Markov Models. In *Proceedings of the 31st Annual Conference of the Canadian Association for Information Science*, 275-287.
[2] Crawford, S.L. et al. 1991. Classification Trees for Information Retrieval. *The 8th International Workshop on Machine Learning*, 245-249.
[3] Lewis, D.D. & Ringuette, M. 1994. A Comparison of Two Learning Algorithms for Text Categorization. In *Third Annual Symposium on Document Analysis and Information Retrieval*, 81-83.
[4] Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *European Conference on Machine Learning, ECML-98*, 137-142.
[5] Joachims, T. 1997. A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 143-151.
[6] Baker, L. D. & McCallum, A. K. 1998. Distributional clustering of words for text classification. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, 96-103.
[7] Larkey, L. S. 1998. Automatic essay grading using text categorization techniques. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, 90-95.
[8] Yang, Y. & Liu, X. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 42-49.
[9] Dumais, S. T., Platt, J., Heckerman, D., & Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proceedings of CIKM-98, 7th ACM International Conference on Information and Knowledge Management*, 148-155.
[10] Li, Y. H. & Jain, A. K. 1998. Classification of text documents. *Computation Journal*. 41, 8, 537-546.
[11] Caropreso, M. F., Matwin, S., & Sebastiani, F. 2001. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In *Text Databases and Document Management: Theory and Practive*, A. G. Chin, ed. Idea Group Publishing Hershey, PA, 78-102.
[12] Sebastiani, F., Sperduti, A., & Valdambrini, N. 2000. An improved boosting algorithms and its application to automated text categorization. In *Proceedings of CIKM-00, 9th ACM International Conference on Information and Knowledge Management*, 78-85.
[13] Mladenic, D. 1998. Feature subset selection in text learning. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, 95-100.
[14] Wiener, E. D., Pedersen, J. O., & Weigend, A. S. 1995. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, 317-332.
[15] Yang, Y. & Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 412-420.
[16] Galaotti, L., Sebastiani, F., & Simi, M. 2000. Experiments on the use of feature selection and negative evidence in automated text categorization. In *Proceedings of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries*, 59-68.
[17] A set of data required for building the statistical model.
[18] A set of data for measuring the system performance.
[19] Online Computer Library Center, Inc (OCLC) is a nonprofit membership organization offering services for libraries and their users. The primary service of this organization is to provide online shared cataloging system for its members.
[20] Proquest Digital Dissertations (PQDD) Hhttp://wwwlib.umi.com/dissertations/H (visited Dec. 2002)