Automated Text Classification Using a Multi-Agent Framework

Yueyu Fu, Weimao Ke and Javed Mostafa
Laboratory of Applied Informatics Research
Indiana University, Bloomington
IN, 47405-3907
(812)856-4182, 01
{yufu, wke, jm}@indiana.edu

ABSTRACT
Automatic text classification is an important operational problem in digital library practice. Most text classification efforts so far concentrated on developing centralized solutions. However, centralized classification approaches often are limited due to constraints on knowledge and computing resources. In addition, centralized approaches are more vulnerable to attacks or system failures and less robust in dealing with them. We present a de-centralized approach and system implementation (named MACCI) for text classification using a multi-agent framework. Experiments are conducted to compare our multi-agent approach with a centralized approach. The results show multi-agent classification can achieve promising classification results while maintaining its other advantages.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *distributed systems, performance evaluation*

General Terms
Algorithms, Design, Experimentation

Keywords
Classification, Multi-Agent System

1. INTRODUCTION
Text classification, broadly defined as determining and assigning topical labels to content, is a fundamental operation in digital libraries. Typically, automatic text classification has been conducted in a centralized architecture [1, 2, & 3]. A single classifier is responsible for classifying all the incoming documents. Theoretically, a centralized classification system, which has all the necessary knowledge and computing resources, can be built to solve any classification problem. However, performing long-term classification may exceed the capabilities of centralized classification systems in practice. For example, the knowledge of a centralized classification system may become "stale" over time. Also, a centralized classification system may be overwhelmed by a large and dynamic document stream (e.g., online news).

The Internet is a distributed system and it offers the opportunity to take advantage of distributed computing paradigms and distributed knowledge resources for classification. With this motivation we attempted to develop a distributed automated classification environment that can offer satisfactory classification performance. Below, we describe the details of the system and experimental studies conducted to evaluate the system on a standard document collection, RCV1-v2 [2].

JCDL '05, June 7–11, 2005, Denver, Colorado, USA

2. MULTI-AGENT CLASSIFICATION
An agent is an autonomous computer program which can emulate certain intelligent behavior and conduct tasks on behalf of its user. Classification agents are independent homogeneous text classifiers except that each agent can only classify content from a limited domain. Agent coordination strategy, which controls agent communication and interaction, is a key component of a multi-agent system. When an agent is unable to identify any or all of the classes for a document, the agent may seek help from other agents based on the agent coordination strategy [4].

Agent representation: Documents and classes are represented in Vector Space Model using TF*IDF term weights. Each agent is represented as a feature vector. For each class, the features are the top ranked terms from the corresponding training documents based on TF*IDF term weights. Each document is represented as a document vector. These terms are the top ranked terms from the whole training set based on the TF*IDF term weights. Cosine similarity score is calculated between a document vector and a class vector. If the similarity score exceeds a pre-defined threshold, the document is considered as a member of this class. Otherwise, this classification fails.

Architecture: Multi-Agent Collaboration and Classification of Information (MACCI), a multi-agent classification system has been implemented using the DIET Agents platform [5]. There are two kinds of agents in MACCI environment: an administration agent which is in charge of distributing the documents from the document pool and a group of classification agents which is responsible for conducting the actual classification tasks.

2.1 Agent Coordination Strategy
A coordination strategy, called multi-agent Good-Neighbor strategy, was developed for MACCI. The hierarchical structure of the classes in the document set RCV1-v2 [2] was utilized to build the Good-Neighbor strategy. There are four parent classes at the top level. Each of these classes has certain number of child classes. Some

child classes have their own child classes. Each agent has two lists of agents, called good neighbors, which can offer classification service: success list and failure list. The success list contains the agents that represent its parent and child classes. The failure list contains the agents that represent the top level parent classes. Below is the coordination algorithm:

1. The administration agent distributes a document from the document pool to a randomly chosen classification agent. This step is repeated after certain interval until the document pool is empty.
2. If an agent successfully classifies a document, it sends the document to the agents in its success list for other potential classification. The help degree is set to 1.
3. If an agent fails to classify a document, it sends the document to agents in its failure list for help. The help degree is set to 1.
4. If an agent successfully classifies a document sent from another classification agent and the help degree is smaller than 4, it sends the document only to the agents that represent its child classes in its success list. The help degree is incremented by 1.
5. If an agent fails to classify a document sent from another classification agent, it doesn't take any action. If none of the agents can classify the document, this case is considered as a NULL classification.

3. EXPERIMENT DESIGN

A set of experiments has been conducted to compare the performance of our multi-agent classification approach with a centralized classification approach. A centralized classification system was developed using the same classification algorithms as the multi-agent classification approach.

3.1 Data set

RCV1-v2 [2] was chosen to be the data set for the experiments. It contains a corpus of more than 800,000 manually categorized newswire stories from Reuters, Ltd. The collection was split into a training set of 23,149 documents and a test set of 781,265 documents. It contains documents from 103 classes. Some examples of classes are corporation, economics, government, and markets. On average, each document is a member of three classes.

3.2 Evaluation Methodology

The classification performance was evaluated using standard effective measures including precision, recall, and F measure. Micro-averaging and macro-averaging methods were used to compute the average F scores. F_1 is defined as follows:

$$F_1 = \frac{2 * \Pr ecision * \mathrm{Re} call}{\Pr ecision + \mathrm{Re} call}$$

4. RESULTS

The experimental results from our centralized and multi-agent approach were collected for comparison. The multi-agent experiment on the entire test set (large) and a small proportion of it (small) are presented below. Lewis and his colleagues conducted experiments on the same document collection using different centralized classification approaches [2]. Their benchmark result using a SVM classifier was 0.82 (micro-averaging F_1) and 0.61 (macro-averaging F_1).

Table 1. Classification effectiveness

Method	Centralized	Multi-agent (small)	Multi-agent (large)
microF1.0	0.72	0.57	0.51
macroF1.0	0.54	0.46	0.40

5. DISCUSSION & CONCLUSION

The micro-averaging F score is always higher than the macro-averaging F score in this case. According to Yang and Liu [1], the micro-averaging score is more influenced by classification performance on common classes and the macro-averaging score is influenced by classification performance on rare classes. Since in the RCV1-v2 training set the common classes have more positive examples than the rare classes, our classifier performed better on the common classes which resulted in higher micro-averaging F score. Overall, the results show the multi-agent approach can provide acceptable classification performance in terms of effectiveness, which can be an alternative when a centralized approach is impossible to realize. In the future, we will experiment with additional coordination strategies such as a market-based strategy. Efficiency should also be measured and further study about how to balance between effectiveness and efficiency is needed.

6. ACKNOWLEDGMENTS

This work was partially supported through a grant from the National Science Foundation Award#:0333623.

7. REFERENCES

[1] Yang, Y., and Liu, X. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 42-49, 1999.

[2] Lewis, D. D.; Yang, Y.; Rose, T.; and Li, F. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, 5:361-397, 2004.

[3] Lewis, D. D. Evaluating and Optimizing Autonomous Text Classification Systems. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 246-254, 1995.

[4] Mukhopadhyay, S., Peng, S., Raje, R., Palakal, M., & Mostafa, J. Large-scale Multi-agent Information Classification Using Dynamic Acquaintance Lists. *Journal of the American Society for Information Science & Technology*, 54(10), 2003.

[5] DIET. DIET Agent Platform, 2004. http://diet-agents.sourceforge.net.