Semi-supervised Text Categorization by Active Search

Zenglin Xu[†] Rong Jin[‡] Kaizhu Huang[†] Michael R. Lyu[†] Irwin King[†]
[†] Dept. of Computer Science and Engineering
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
{zlxu,kzhuang,lyu,king}@cse.cuhk.edu.hk
[‡] Dept. of Computer Science and Engineering
Michigan State University
East Lansing, MI, 48824
rongjin@cse.msu.edu

ABSTRACT

In automated text categorization, given a small number of labeled documents, it is very challenging, if not impossible, to build a reliable classifier that is able to achieve high classification accuracy. To address this problem, a novel web-assisted text categorization framework is proposed in this paper. Important keywords are first automatically identified from the available labeled documents to form the queries. Search engines are then utilized to retrieve from the Web a multitude of relevant documents, which are then exploited by a semi-supervised framework. To our best knowledge, this work is the first study of this kind. Extensive experimental study shows the encouraging results of the proposed text categorization framework: using Google as the web search engine, the proposed framework is able to reduce the classification error by 30% when compared with the state-of-the-art supervised text categorization method.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Search and Retrieval; I.5.2 [**Design Methodology**]: Classifier Design and Evaluation

General Terms

Algorithm, Performance, Experimentation

Keywords

Text Categorization, Semi-supervised Learning

1. INTRODUCTION

The goal of automated text categorization is to automatically classify documents into predefined categories. Since the performance of supervised statistical classifiers often depends on the availability of labeled examples, one of the major bottlenecks toward automated text categorization is to collect sufficient numbers of labeled documents because of the high cost in manually labeling documents. One way to address the problem of small-size sample is to exploit the unlabeled documents by so-called semi-supervised learning methods [1]. However, in order to exploit the semi-supervised learning techniques, one of the major issues is to obtain a multitude of unlabeled documents that are relevant to the target categories.

One way to collect the unlabeled documents is through the web search engines. In order to retrieve web documents that are relevant to the target topics, we will first identify the keywords from a few labeled documents that are closely related to the target topics. Web documents will then be retrieved by the web search engine using the textual queries that are constructed based on the identified keywords. Finally, the retrieved web documents will be combined with the labeled documents to construct a text classification model using the semi-supervised learning techniques. We refer to this framework as "**Semi-supervised Text Categorization by Active Search**", whose goal is to enhance a text classification model by actively exploiting the unlabeled web documents via the web search engines. The key features of this framework include (1) *Query generation* that generates the textual queries for document retrieval by analyzing the content of labeled documents, (2) *Document retrieval* that retrieves the web documents through the web search engine by using the generated queries, and (3) *Semi-supervised text categorization* that constructs text classification models by utilizing both the labeled documents and the unlabeled web documents which are retrieved by the web search engine. To our best knowledge, this work is the first study on the subject of actively retrieving related documents from the Web as a complementary information source for supervised text categorization.

2. PROPOSED FRAMEWORK

We focus on the stages of query generation and semi-supervised text categorization as document retrieval is naturally cared by search engines.

The objective of query generation is to construct queries that are likely to retrieve documents relevant to the target categories. A straightforward approach is to construct queries by selecting informative keywords using statistical measures such as TF, TF/IDF, information gain, and χ^2. However, the conventional metric may not find the most representative words due to the "**sparse data**" problem. Moreover, the selected query words may be unrelated when extracted from different documents, resulting irrelevant document retrieved by web search engine. We refer to it as the "**unrelated query words**" problem.

To address the two problems, we propose to generate a query for every labeled document. Let $\mathbf{x}_i$ be a document assigned to the category c_i. Let V_i denote the vocabulary used by document $\mathbf{x}_i$, i.e., $V_i = \{k|x_{i,k} > 0\}$. To generate

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.
ACM 978-1-59593-991-3/08/10.

a query $\mathbf{q}_i$ for category c_i based on document $\mathbf{x}_i$, we first restrict the words used by query $\mathbf{q}_i$ to vocabulary V_i, i.e., $q_{i,k} = 0$ if $k \notin V_i$. To measure the informativeness of words, we introduce a non-negative weight $w_i \geq 0$ for each word: the more informative a word is, the larger the weight will be. We relabel each document $\mathbf{x}_j$: $y_j = +1$ if $\mathbf{x}_j$ is assigned to category c_i, and $y_j = -1$ otherwise. We follow the framework of Support Vector Machine (SVM) [3] and determine the word weights by the following optimization problem:

$$\min_{\mathbf{w},\xi} \quad \sum_{j \in V_i} w_j + C \sum_{k=1}^{n_l} \xi_k \tag{1}$$

$$\text{s. t.} \quad y_k \left(\sum_{j \in V_i} w_j x_{k,j} + b \right) \geq 1 - \xi_k, \xi_k \geq 0, k = 1, \ldots, n_l \ ,$$
$$w_j \geq 0, \ \forall j \ , \quad w_j = 0, \ \forall j \notin V_i.$$

where C is the parameter that weights between the classification error $\sum_{k=1}^{n_l} \xi_k$ and the regularization term $\sum_{j \in V_i} w_j$, and b is the bias. The statistical analysis of the above model can be found in the long version of this paper. Finally, the query $\mathbf{q}_i$ is formed by choosing the first k words with the largest values of $\mathbf{w}$. We denote by $\mathcal{U}^{(i)}$ the retrieved documents, and refer to the proposed method as "**discriminative query generation**".

In addition to query generation, we present two different approaches for text categorization that combine both the labeled documents $\mathcal{D}$ and the unlabeled retrieved documents $\mathcal{U}$: (1) the *auxiliary approach* that assumes all the retrieved documents in $\mathcal{U}^{(i)}$ belong to the category c_i, and (2) the *semi-supervised learning approach* that does not assume any relationship between the class labels for retrieved documents $\mathcal{U}^{(i)}$ and the class label c_i.

In the auxiliary approach, we will train a classification model using both the labeled documents and the auxiliary labeled documents. We thus have the following concrete optimization problem for text categorization:

$$\arg\min_{\mathbf{w},b} \quad \lambda \|\mathbf{w}\|_2^2 + \sum_{\mathbf{x}_i \in \mathcal{D}} \xi_i + \gamma \sum_{\mathbf{x}_j \in \mathcal{U}} \xi_j \tag{2}$$

$$\text{s. t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i \ \mathbf{x}_i \in \mathcal{D} \ ,$$
$$y_j^*(\mathbf{w}^\top \mathbf{x}_j + b) \geq 1 - \xi_j, \quad \forall j \ \mathbf{x}_j \in \mathcal{U} \ ,$$

where γ and λ are trade-off parameters determined manually.

Different from the auxiliary approach, the semi-supervised learning approach optimizes not only the classification function $h(\mathbf{x})$, but also the class label assigned to the unlabeled data $\mathbf{y}^*$. The related optimization problem is formulated as follows:

$$\arg\min_{\mathbf{w},b,\mathbf{y}^*} \quad \lambda \|\mathbf{w}\|_2^2 + \sum_{\mathbf{x}_i \in \mathcal{D}} \xi_i + \gamma \sum_{\mathbf{x}_j \in \mathcal{U}} \xi_j \ , \tag{3}$$

$$\text{s. t.} \quad y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i \ \mathbf{x}_i \in \mathcal{D} \ ,$$
$$y_j^*(\mathbf{w}^\top \mathbf{x}_j + b) \geq 1 - \xi_j, \quad \forall j \ \mathbf{x}_j \in \mathcal{U} \ ,$$

which can be efficiently solved by CCCP [2].

3. EXPERIMENT

Nine subsets of text documents are selected from three benchmark text collections, including 20 Newsgroups, Reuters-21578, and Ohsumed. We randomly select 5 labeled documents per class to form the training set for each data set, and use the remaining documents as the test set. For each document, one query of three terms is generated to retrieve similar documents from the Web. We use google as the search engine. We download the first 100 documents returned by each query based on the assumption that most search engines rank relevant documents before the irrelevant ones. We employ SVM as our baseline algorithm. Two learning algorithms that can utilize the unlabeled documents are implemented for text categorization, including the auxiliary SVM (abbreviated as Auxi-SVM) and semi-supervised SVM (abbreviated as Semi-SVM).

Table 1: The classification accuracy (%) of text categorization

Data set	SVM	Auxi-SVM	Semi-SVM
male vs. female	47.6	**76.1**	73.1
bacterial vs. virus	61.8	77.6	**78.3**
musculo vs. digestive	69.9	71.3	**77.0**
fourDisease	31.6	38.4	**58.0**
ship vs. trade	94.1	95.5	**95.9**
corn vs. wheat	69.2	69.0	**71.6**
money vs. trade	80.6	88.8	**88.9**
auto vs. motor	59.4	69.1	**69.2**
sci	35.5	56.1	**56.8**
average	61.1	71.3	**74.3**

Table 1 summarizes the classification accuracy of SVM and the proposed methods. Both the categorization methods (Auxi-SVM and Semi-SVM) using the retrieved documents achieve significantly higher accuracy than the supervised method for almost all data sets. The overall error reduction over SVM is 26.3% for Auxi-SVM and 34.0% for Semi-SVM. In conclusion, using the documents retrieved from the Internet can greatly improve the classification accuracy. The advantage is more manifest when the semi-supervised SVM is adopted for text categorization.

4. CONCLUSIONS

In this paper, we presented a general framework for semi-supervised text categorization that collects the unlabeled documents via web search engines and utilizes them to improve the accuracy of supervised text categorization. Extensive experiments have demonstrated that the proposed semi-supervised text categorization framework can significantly improve the classification accuracy.

5. ACKNOWLEDGMENTS

The work was substantially supported by a grant from the National Science Foundation, U.S. (IIS-0643494) and two grants from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK4150/07E and Project No. CUHK4125/07).

6. REFERENCES

[1] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[2] R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale transductive SVMs. *JMLR*, 7:1687–1712, 2006.

[3] V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.