Automatic Subject Heading Assignment for Online Government Publications Using A Semi-supervised Machine Learning Approach

Xiao Hu[1]

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street., Champaign, IL, 61820. Email: xiaohu@uiuc.edu

Larry S. Jackson

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street., Champaign, IL, 61820. Email: lsjackso@uiuc.edu

Sai Deng

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street., Champaign, IL, 61820. Email: saideng@uiuc.edu

Jing Zhang

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, 501 E. Daniel Street., Champaign, IL, 61820. Email: zhang20@uiuc.edu

As the dramatic expansion of online publications continues, state libraries urgently need effective tools to organize and archive the huge number of government documents published online. Automatic text categorization techniques can be applied to classify documents approximately, given a sufficient number of labeled training examples. However, obtaining training labels is very expensive, requiring a lot of manual labor. We present a real world online government information preservation project (PEP[1]) in the State of Illinois, and a semi-supervised machine learning approach, an Expectation-Maximization (EM) algorithm-based text classifier, which is applied to automatically assign subject headings to documents harvested in the PEP project. The EM classifier makes use of easily obtained unlabeled documents and thus reduces the demand for labeled training examples. This paper describes both the context and the procedure of such an application. Experiment results are reported and other alternative approaches are also discussed.

The Government Documents Problem Space

As the dramatic expansion of online publications continues, government organizations providing information to the public urgently need effective tools to organize and archive documents published online. The Preserving Electronic Publications (PEP) project, a cooperation of the Illinois State Library (ISL) and the Graduate School of Library and Information Science (GSLIS) in University of Illinois at Urbana-Champaign, is aiming to effectively organize and archive a huge number of government documents within or beyond Illinois electronically published on the Web. In order to achieve this goal, one important task is organizing documents as hierarchies of concepts. In this regard, a consortium of State Libraries and Archives adapted the Global Information Locator Services (GILS) tag set and published a customized subject heading list, also called a "Topic Tree" [2] due to its hierarchical structure, facilitating information organization in State libraries and government agencies (Zussy 2000). However, manually assigning these subject headings to the huge amount of online publications is too time-consuming and expensive. Automatic methods can perhaps be used to help generate subject headings based on an analysis of the text of the files, and a predefined set of subject headings. In this poster, we describe a set of experiments

[1] Please send correspondence to the first author.

applying a semi-supervised machine learning approach, an EM algorithm-based text classifier, to online government publications harvested in the PEP project.

Datasets and Systems

The datasets and systems involved in the PEP project provide important practical contexts for this research.

CEP web harvesting, metadata extraction and retention

Capturing Electronic Publications (CEP) is one of ISL system initiatives. The goal is to support multi-state deployment of our freeware-based web archiving facility. Currently, it not only covers 225 government websites in Illinois, but also harvests government information for other six states (NC, MT, AZ, AK, UT, WI). Every month, the spider automatically harvests all electronic documents published on the websites of various governmental agencies. So far, there are more than 440,000 web documents for Illinois alone. However, among the large number of documents, only very few of them provide author-generated metadata, of which the quantity and quality are far from desirable (Jackson 2003). To solve the problem, metadata inference based on data mining and information retrieval techniques might be used to compensate for the lack of author-generated metadata.

EDI Depository

The Electronic Documents Initiative (EDI)[3] provides permanent retention and web access to "official" State publications existing in electronic format. The e-documents are defined, gathered and retained in response to human decisions. To better support metadata generation, the authoring agency specifies cataloging metadata via ISL's Metadata Generator website. As a result of these human efforts, there is a much greater degree of author-generated metadata in the EDI depository than in the Illinois State Government web as a whole.

IGI Search engine

Both CEP and EDI collections are integrated with Illinois Government Information (IGI) search engine[4], to provide users with full access to online government information. Searching functions will be greatly enhanced with rich metadata. For example, currently the IGI search engine supports searching by subject, website, originator, etc. However, as mentioned above, author-generated metadata are very rare. To solve the problem, we've applied text mining techniques to automatically extract metadata such as keywords, description, language, etc. The semi-supervised learning approach presented in this poster is used to automatically generate subject heading metadata.

The Semi-Supervised Approach

There are two kinds of learning approaches in machine learning literature: supervised and unsupervised learning (Mitchell 1997). Completely unsupervised methods have difficulty in labeling the relations between documents and pre-defined subject headings since no labeled documents are available as training examples. Supervised methods are constrained by the often small number of labeled documents available for use in training. Semi-supervised methods can potentially help assign subject headings, while reducing the number of labeled documents below that required by supervised methods. The approach adopted in this work is based on an EM (Expectation - Maximization) algorithm which exploits easily obtained unlabeled documents by assuming the unlabeled documents and labeled ones share the same probabilistic distribution (Nigam et al. 2000). The EM algorithm is a general method of finding the maximum-likelihood estimate of an underlying distribution from a given data set when the data is incomplete or has missing values. In the case of subject heading assignment, the missing values are the subject headings of unlabeled documents. After the underlying distribution has been estimated by the EM algorithm, subject headings of new documents will be assigned according to such distribution.

The EM Algorithm

With a limited number of labeled documents, the EM algorithm can help augment the training data set by incorporating unlabeled documents. It iterates alternating back and forth between two steps. Beginning with an initial probabilistic distribution of the documents, the E (Expectation)-step estimates the expectation of the missing values: subject headings of unlabeled documents. The M (Maximization)-step uses those labels we just got in the E-step, together with labeled training documents, to re-estimate the document distribution. Then the E-step is running again with the updated document distribution. While the two steps are repeated, it is guaranteed that the estimation will hit optimal values, which is called "convergence" (Dempster et al. 1977). As this is a semi-

supervised approach, the originally available labeled training documents are used to set the initial probabilistic distribution, i.e. to get a reasonable start point of the iteration. This is an advantage of semi-supervised approaches over completely unsupervised ones.

Experiments and Results

In order to verify the effectiveness of the EM based semi-supervised learning approach, we compare it with a Naïve Bayesian (NB) text classifier which is a supervised learning approach and has been widely used in text categorization literature because of its efficient nature (Sebastiani 2002, Yang 1999). Moreover, because Naïve Bayes shares the same probabilistic foundation with the EM algorithm, such comparison can focuses on the advantages of semi-supervised learning approaches over supervised ones. Specifically, we want to see 1) with the same amount of labeled documents, can the EM approach achieve better results? 2) to achieve the same result, does the EM approach require less labeled documents?

Data Sets

Our cataloguer (a graduate student in the library school) manually labeled 194 documents as training data. This set of training documents, and various sets of unlabeled documents are randomly sampled from CEP Illinois collection covering a variety of agencies. The testing documents are selected from those with author-provided subject headings and are rectified by the cataloguer. The *Policies for Assigning Subject Headings of Library of Congress Subject Headings* (LCSH 1990) is adapted as the cataloging guideline. As an output, 604 subject headings on the top two levels of the topic tree are assigned to the training data set whereas 410 to testing set. The document distribution across subject headings is very uneven, and some headings are too scarce to find a corresponding document in the sampled collection. Therefore we keep headings with at least one labeled document and remove others from the topic tree. This results in 21 headings on the top level and 180 on the second.

Data Preprocessing

Common short words are removed from the collection. Morphological analysis is employed to replace numbers, year, month and weekdays with a single token respectively. After the terms are stemmed, there are 4,597 unique terms in labeled training documents. Finally 1,500 features are selected by the Information Gain criterion.

Evaluation Measures

Precision (P) and recall (R) are the most accepted measures in information retrieval and text categorization. Precision is the proportion of relevant ones among all retrieved documents, while recall is the proportion of retrieved ones among all relevant documents. F measures balance precision and recall into one single value (Van Rijsbergen 1979), among which F1 measure gives equal weight to precision and recall. For the task of subject heading assignment, precision and recall are equally important, so we adopt F1 measure in our experiments. In text categorization tasks with more than two categories, one needs to average measures over categories. There are two ways to do so: one is to compute evaluation measures for each category, then average the measures over all categories, which is called macro-averaging; the other is micro-averaging, which is to aggregate the numbers of retrieved documents and relevant documents over all categories, and then calculate evaluation measures based on the aggregated numbers. Macro-averaging gives equal weight to each category whereas micro-averaging gives equal weight to each document. It is not hard to see that micro-averaging can be dominated by categories with more documents (Manning & Schütze 1999). For this reason, our result report emphasizes on macro-averaging values.

Results

Figure 1 shows macro-averaging performances with regard to various numbers of unlabeled documents. Improvements are observed when the number of unlabeled documents ranges from 1600 to 2400. The benefit comes from the unlabeled documents. With so few labeled training examples (194 documents), the Naïve Bayesian classifier cannot learn accurate estimates. On the other hand, the EM approach improves the estimates by taking into account unlabeled documents of about 10 times as many as the labeled documents. It should be noticed that the performances drop when still more unlabeled documents are involved. That may be caused by the violation of the assumption that labeled and unlabeled documents share the same distribution. When the

number of unlabeled documents becomes large, chances are the limited number of labeled documents cannot represent the distribution of so many unlabeled documents.

Figure 2 compares the learning curves of the EM classifier and the Naïve Bayesian (NB) classifier. When the number of labeled documents is small, the macro F1 values have not clearly demonstrated the benefit of the EM classifier until all the 194 labeled documents are used. We expect the trend would be more apparent if given more labeled documents. However, the merit of the EM classifier is shown earlier by the measure of micro F1. In order to achieve the performance of 0.28 in micro F1, the Naïve Bayesian classifier needs 150 labeled documents, whereas only 100 are needed for the EM classifier, which is 33% more efficient in terms of using labeled documents.

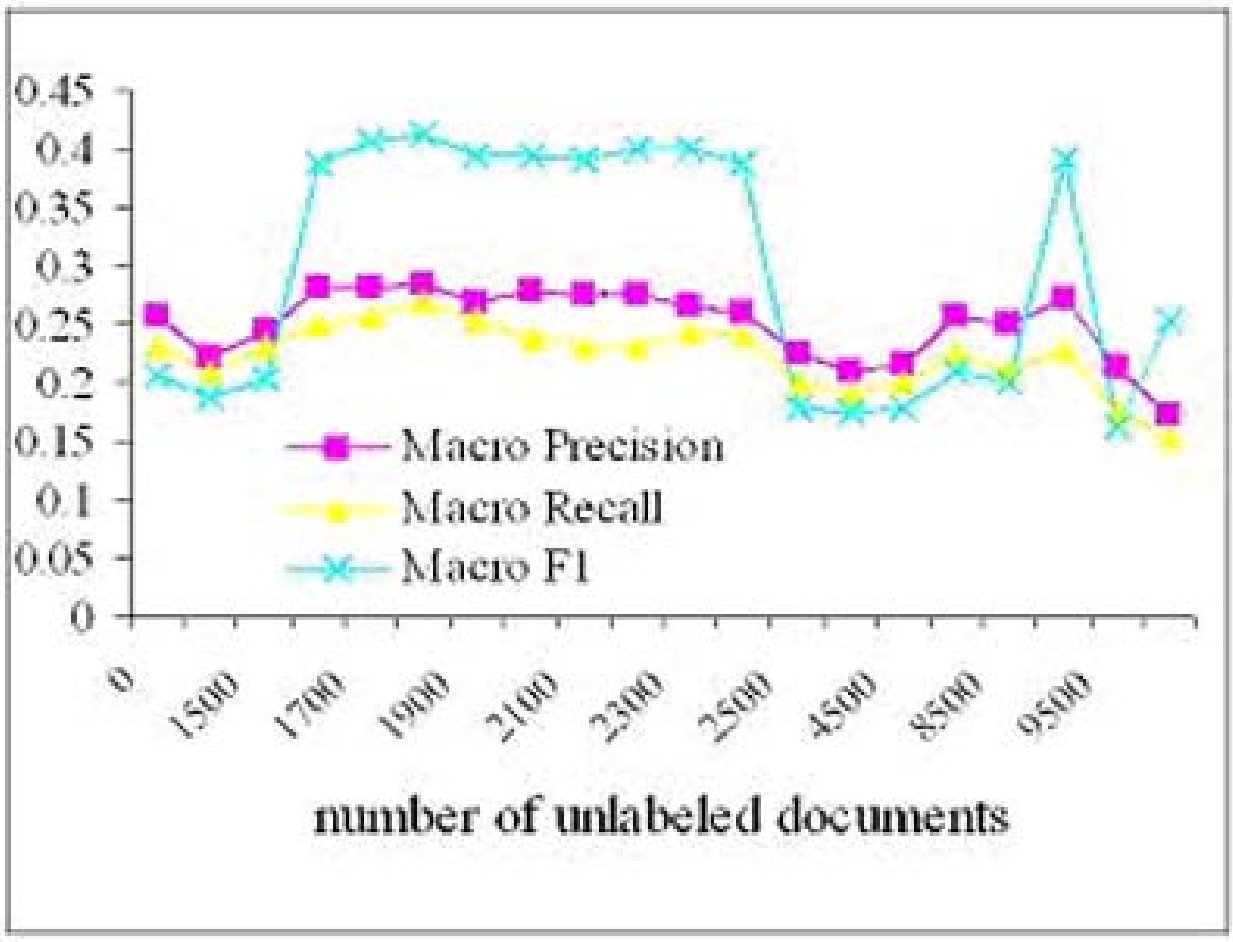

Figure 1: Macro averaging performance. The number 0 on the x-axis corresponds to the Naïve Bayesian classifier which doesn't need unlabeled documents.

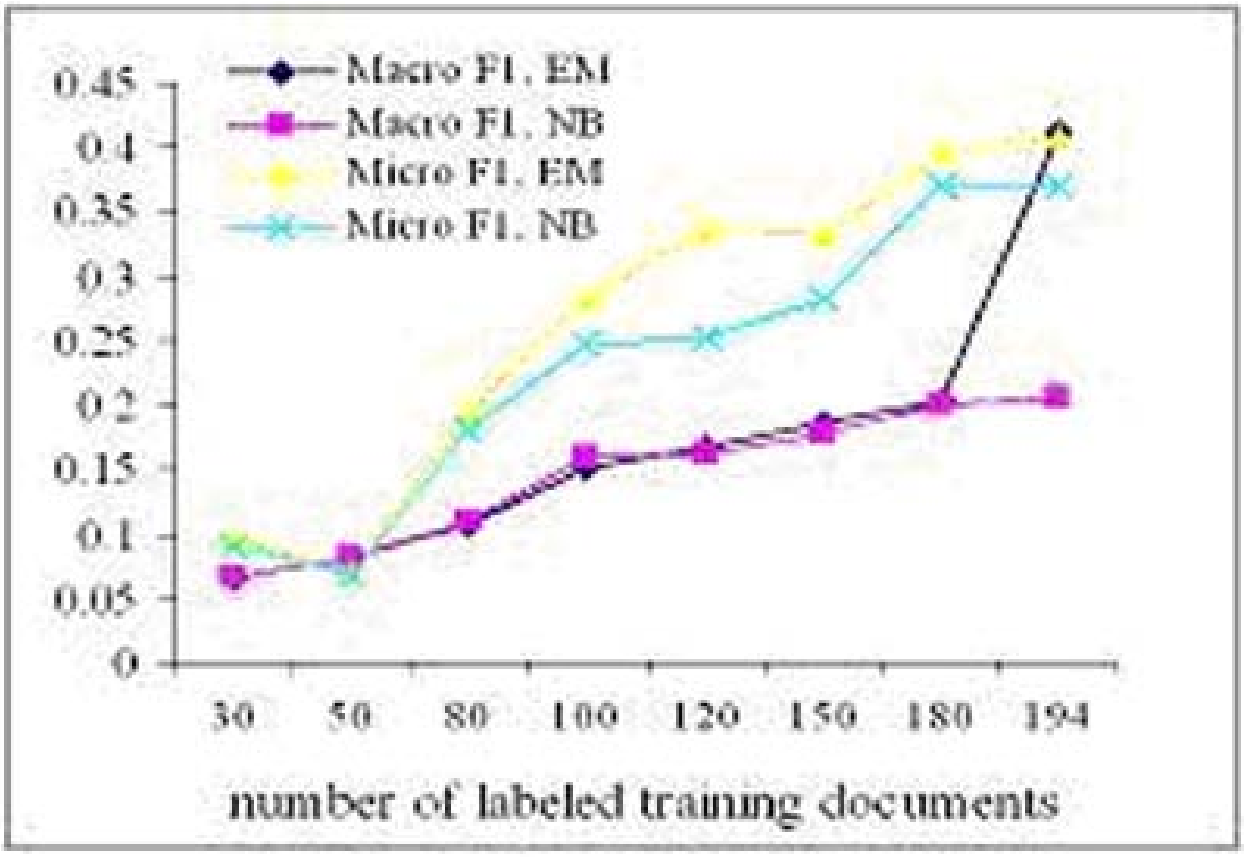

Figure 2. Learning curves of the EM and NB classifiers. In the EM classifier, the numbers of unlabeled documents are set as 10 times more than corresponding labeled ones.

Discussion on Other Techniques

Semi-supervised learning can save a lot of work on labeling training examples, but it is by no means the only approach to automatic subject heading assignment. In this section, we describe some other approaches we are using to solve the problem.

K-Means

K-means clustering method is one of the most popular methods in unsupervised learning community. It takes the desired number of groups, k, and partitions a set of objects into k clusters so that the resulting intra-cluster similarity is high but the inter-cluster similarity is low. In the PEP project, we tried the SimpleKMeans clustering method, which is a centroid based clustering method and implemented by the Weka project (Witten & Frank 2005). It achieved 32% precision on the whole Illinois collection of 422,152 documents. However, the result cannot be directly compared to that of the EM algorithm because EM, as a semi-supervised approach, was limited by the number of available labeled training examples. With more labeled examples, EM algorithm can be expected to achieve better results. On the other hand, we are also experimenting other refined clustering methods, some of which yield better precisions than K-means.

Collection-level default subject headings

Another resort we are pursuing is to assign subject headings to documents harvested from each website as a whole, according to the functions of the agency who owns the website. Cataloging a website as a whole seems more efficient than cataloging every document on this website. However, the approach will result in coarse granularity of assigned subject headings, and it is unclear whether such coarse granularity will facilitate or hinder users' searching and access to online government information. Therefore, we believe user studies are needed to compare the approaches to one another.

Conclusions and Future Work

This paper gives an example of applying a semi-supervised text categorization approach in a real-word practice of archiving and searching online government publications. The experiments provide a reference to other projects working with online government information. The semi-supervised approach works towards reducing manual work in subject heading assignments. Without considerable reduction in the cost of assignment of subject headings, it is highly unlikely the existing inventory of electronic publications, particularly those very many documents already published on the web, will ever be so processed. Accordingly, searches and browsing involving restrictions by subject heading will continue to be impaired. Progress in this topic will greatly improve the efficiency of electronic information search and retrieval.

In future work, we will compare this approach with others regarding to effectiveness as well as efficiency. Unlike pure laboratory work, in such a practical project as PEP, efficiency of resource consumption (e.g. response time, CPU complexity, etc) is an important criterion of assessment. In addition, our ultimate goal is to better serve users' information needs, therefore we will perform a formal user needs assessment in the near future. Finally, after thorough testing, the approaches will be deployed in Illinois' IGI search engine, to be used and benefit real users.

ACKNOWLEDGMENTS

This work was sponsored in part by a National Leadership Grant from the Institute of Museum and Library Services and by the Illinois State Library.

REFERENCES

A.P.Dempster, N.M.Laird, & D.B.Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, 39(B)

Jackson, L. S. (2003). Preserving State Government Web Publications -- First-Year Experiences. In Proceedings of the National Conference on Digital Government Research dg.o2003. Digital Government Research Center, Marina del Rey, CA. 2003. pp. 109-114.

LCSH (1990). Library of Congress Subject Headings: Principles of Structure and Policies for Application, Annotated Version, prepared by Lois Mai Chan for the Library of Congress; Cataloging Distribution Service, Library of Congress, Washington, D.C.

C. Manning & H. Schütze (1999). Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA, 1999

T. Mitchell (1997). Machine Learning, McGraw Hill

K. Nigam, A.McCallum, S. Thrun & T. Mitchell (2000). Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning, 39(2/3). pp. 103-134

C. J. Van Rijsbergen (1979). Information Retrieval, Butterworths,

F. Sebastiani (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1) pp.1-47

I. Witten & E. Frank (2005). Data Mining: Practical Machine Learning Tools and Techniques". Morgan Kaufmann, San Francisco, 2005

Y. Yang (1999). An evaluation of statistical approaches to text categorization. Journal of Information Retrieval v(1) pp. 69-90

N. Zussy (2000). The Birth and Development of Find-It!: Washington State's Government Information Locator Service", in First Monday, volume 5, number 6 (June 2000) pp.9 http://www.firstmonday.dk/issues/issue5_6/zussy/index.html

[1] http://www.isrl.uiuc.edu/pep/

[2] http://www.finditillinois.org/metadata/subjtree.pdf

[3] http://iledi.org

[4] http://findit.lis.uiuc.edu/