Synopsis Information Extraction in Documents Through Probabilistic Text Classifiers

Jantima Polpinij[1] and Aditya Ghose[2]

[1] Faculty of Informatics, Mahasarakham University, Mahasarakham 44150 Thailand
jantima.p@msu.ac.th
[2] School of Computer Science and Software Engineering, Faculty of Informatics,
University of Wollonong, Wollongong, 2500 NSW, Australia
aditya@uow.edu.au

1 Introduction

Digital Libraries currently use several advanced information technologies to organize information and make it easy accessible to users. Current digital library trends to be dynamic digital library [1]. It is possible that business rules also can be approached for improving dynamic digital library. Business rules [2] are statements that define or contain some aspects of IT systems by providing a foundation for understanding how an IT system functions. At present, the need for automated business rules is becoming more essential because of the increasing usage of IT systems. However, it is not easy to extract business rules because they are written in a natural language structure and much of it is ignored. Therefore, one important question in this research area is how to automatically extract a business rule from a document? Based on this, information extraction (IE) [3] typically can be applied. Basically, IE is to transform text into information that is more readily analyzed. We believe that if the content of a document is decreased, the accuracy of rules extraction may be increased logically. With this assumption, if irrelevant information is filtered from the document, it is possible to easily extract business rules from the rest. Therefore, this research proposes a method based on probabilistic text classifier to extract synopsis information. It could be said that this work is the pre-processing of a business rules extraction methodology.

2 Research Methodology and Results

Before learning method, a text document collection must be transformed into a representation which is suitable for computation. The ordinary way of document representation is usually as a structured "bag of words" [4]. Then it will contain each unique word that becomes a feature, including the number of times the word occurs in the document. In addition, we applied the Chi-squared technique (χ^2) [5] to reduce feature size. Afterwards, finding term word weighting is critical in term-based retrieval since the rank of a document is determined by the weights of the terms. We certainly use the popular term weighting for our work. It is called *TF-IDF* [5].

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 508–509, 2007.

For learning task, the Naïve Bayes [6] classifier is applied. It uses a set of training documents to estimate parameters, and then use the estimated model to filter information in documents. Suppose that we have documents $D = \{d_1,...,d_{|D|}\}$, where Φ is parameter and we use the notation $c_j \in C = \{c_1,...,c_{|c|}\}$ to indicate both j-th component and j-th class. We assume that the documents are generated by a mixture model and there is a one-to-one correspondence between the class labels and the mixture component. Each document d_i is generated by choosing a mixture component with the class prior probabilities $P(c_j;\ \Phi)$, and having this mixture component generate a document according to its own parameters, with distribution $P(d_i \mid c_j;\ \Phi)$. So, we can identify the likelihood of a document as a sum of total probability over all generative components:

$$P(d_i \mid \Phi) = \sum_{j=1}^{|C|} P(c_j \mid \Phi) P(d_i \mid c_j; \Phi)$$

To this end, we have to choose $\text{argmax}_j\ P(c_j| d_i;\ \Phi)$ as the best class of the document. In this research, we also choose the probability that is better of the two probabilities: the "*relevant*" and "*irrelevant*" classes. Finally, the "*relevant*" is information that is used in the process of business rule extraction.

We used only the misc.forsale topic of the 20 newsgroup dataset for our work. We randomly selected 700 documents for training and 300 documents for testing. After the classifier model finished, we evaluated the results of the experiments by using *F-measure* [4]. We used the classifier model to filter irrelevant sentences in each document. If some sentences in a document are in irrelevant classes, they are filtered from the document. The model shows accuracy at 77%. As a result, it demonstrates that this method can provide more effectiveness for filtering irrelevant information from a document.

References

[1] Walker, A.: The Internet Knowledge Manager, Dynamic Digital Libraries, and Agents You Can Understand, D-Lib Magazine, http://www.dlib.org/dlib/march98/walker/03walker.html

[2] Zsifkov, N., Campeanu, R.: Information technology: Business rules domains and business rules modeling. In: Proceedings of the 2004 international symposium on Information and communication technologies (ISICT) (2004)

[3] Turmo, J., Ageno, A., Catalá, N.: Adaptive information extraction, ACM Computing Surveys (CSUR). ACM Press, New York (2006)

[4] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. The ACM Press, New York (1999)

[5] Yang, Y., Pederson, J.O.: A Comparative Study on Features selection in Text Categorization. In: Proceedings of the 14th international conference on Machine Learning (ICML), Nashville, Tennessee, pp. 412–420 (1997)

[6] Nigam, K., Maccallum, A.K., Thrun, S., Mitchell, T.: Text Classification from Labeled and Unlabeled Document using EM. Machine Learning 39(2/3), 103–134 (2000)