Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

Automatic construction of ontology from text databases

N. Zhong[1], Y.Y. Yao[2] & Y. Kakemoto[3]
[1] *Department of CSSE, Faculty of Engineering, Yamaguchi University, Japan.*
[2] *Department of Computer Science, University of Regina, Canada.*
[3] *The Japan Research Institute, Japan.*

Abstract

The paper describes a multi-phase process of automatic construction of the *domain specific* ontology from text databases, in which various text mining and natural-language understanding methods are used. The ontology we wish to develop describes a well-defined technical domain. Hence it is called a *domain specific* ontology, opposed to *universal* ontologies. We discuss the major techniques used in the process and show some preliminary results.

1 Introduction

It has been recently recognized in the KDD (Knowledge Discovery and Data Mining) community that mining from semi-structured or unstructured text (text mining for short) is an increasingly important research topic, and there are enormous potential applications [1]. Much of data is now in textual form. This could be data on the world wide web, e-mails, library, or electronic papers and books, among others, namely *text databases* in this paper. Information discovery from Internet and electronic commerce are some of the potential applications [6].

Although many techniques and systems for knowledge discovery from relational databases have been developed, few of them can be directly applied to text data. Text mining is a much more complex task as it involves dealing with inherently unstructured and fuzzy data. Text mining is a multidisciplinary field, involving various techniques such as data mining, in-

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

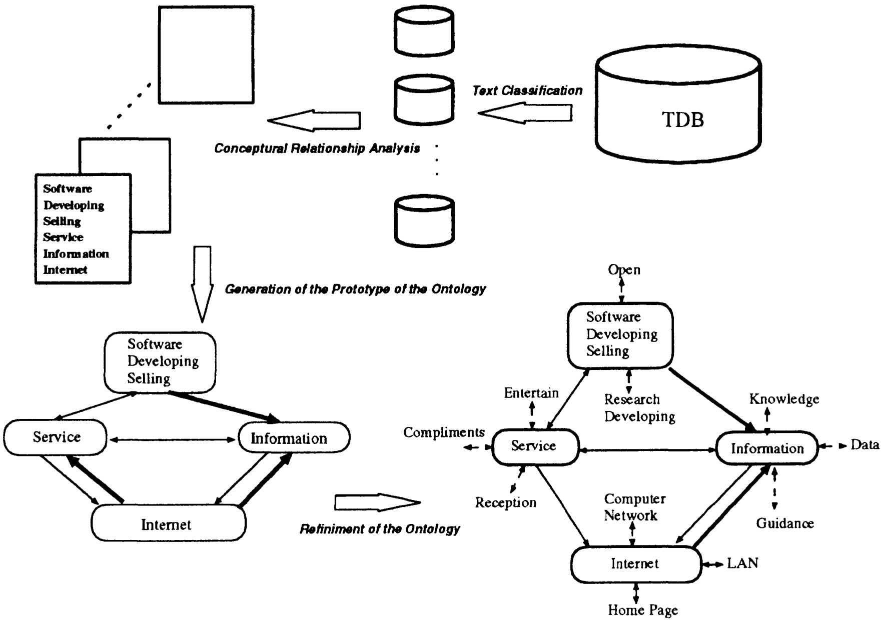

Figure 1: A sample process of construction of the ontology

formation retrieval, natural-language understanding, case-based reasoning, statistics, and intelligent agent technology.

Construction of ontology from technical texts is one of important tasks in text mining [8, 11]. The ontology we wish to develop describes a well-defined technical domain. Hence it is called a *domain specific* ontology [1], opposed to *universal* ontologies. The process of construction of the ontology is a multi-phase process in which various text mining and natural-language understanding methods are used.

This paper describes a process of automatic construction of the *domain specific* ontology from text databases. Figure 1 shows a sample process of construction of the ontology on software marketing. The major steps in the process include morphological analysis, text classification, generation of classification rules, conceptual relationship analysis, generation of ontology, as well as refinement and management of ontology. A thesaurus is necessary to be used as a background knowledge base in the process. We stress that the process is iterative, and may repeat at different intervals when new/updated data come. We have already finished several parts of the proposed system and are in the process of extending the system to include several more

[1]The *domain specific* ontology is also called the *task* ontology.

capabilities of text mining and natural-language understanding. We discuss the major techniques used in the process and show some preliminary results.

2 Text classification

In order to discover a *domain specific* ontology from text databases, we first need to annotate the texts with class labels. This annotation task is that of text classification. However, it is expensive that the large amounts of texts are manually labeled. This section introduces a semi-automatic approach to classify text databases, which is based on uncertainty sampling and probabilistic classifier. The main contribution of ours is to extend the method proposed by Lewis et al. [10] for multiple classes classification.

We use a variant of the Bayes' rule below:

$$P(C|w) = \frac{\exp(\log \frac{P(C)}{1-P(C)} + \sum_{i=1}^{d} \log(P(w_i|C)/P(w_i|\overline{C})))}{1+\exp(\log \frac{P(C)}{1-P(C)} + \sum_{i=1}^{d} \log(P(w_i|C)/P(w_i|\overline{C})))} \tag{1}$$

where $w = \{w_1, \ldots, w_d\}$ is a set of the terms in a text, and C is a class. Although we treat, in this equation, only two classes $C_1 = C$ and $C_2 = \overline{C}$ with $P(\overline{C}) = 1 - P(C)$, it can be extended to deal with multiple classes classification by using the method to be stated in the end of this section.

However Eq. (1) is rarely used directly in text classification, probably because its estimates of $P(C|w)$ are systematically inaccurate. Hence we use Logistic regression, which is a general technique for combining multiple predictor values to estimate a posterior probability, in Eq. (1). Thus, we obtain the following equation:

$$P(C|w) = \frac{\exp(a + b\sum_{i=1}^{d} \log(P(w_i|C)/P(w_i|\overline{C})))}{1+\exp(a + b\sum_{i=1}^{d} \log(P(w_i|C)/P(w_i|\overline{C})))}. \tag{2}$$

Intuitively, we could hope that the logistic parameter a would substitute for the hard-to-estimate prior log odds in Eq. (1), while b would serve to dampen extreme log likelihood ratios resulting from independence violations.

Furthermore, we use the following equation to estimate the values $P(w_i|C)P(w_i|\overline{C})/P(w_i|C)P(w_i|\overline{C})$ as the first step in using Eq. (2),

$$\frac{P(w_i|C)}{P(w_i|\overline{C})} = \frac{\frac{c_{pi}+(N_p+0.5)/(N_p+N_n+1)}{N_p+d(N_p+0.5)/(N_p+N_n+1)}}{\frac{c_{ni}+(N_n+0.5)/(N_p+N_n+1)}{N_n+d(N_n+0.5)/(N_p+N_n+1)}} \tag{3}$$

where N_p and N_n are the numbers of terms in the positive and negative training sets, respectively, c_{pi} and c_{ni} are correspondingly the numbers of examples of w_i in the positive and negative training sets, respectively, and d is the number of different terms in a text.

Based on the preparation stated above, we briefly describe the main steps of text classification below:

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

Step 1. Select examples (terms) as an initial classifier for N classes by a user and all the N classes are regarded as a set of the negative classes.

Step 2. Select a class from the set of the negative classes as a positive class, and the remaining ones are regarded as a set of the negative classes.

Step 3. While a user is willing to label texts.

Step 3.1 Apply the current classifier to each unlabeled text.

Step 3.2 Find the k texts for which the classifier is least certain of class membership by computing their posterior probabilities in Eq (2).

Step 3.3 Have the user label the subsample of k texts.

Step 3.4 Train a new classifier on all labeled texts.

Step 4. Repeat *Step 2.* to *Step 3.* until all classes were selected as a positive class.

Selecting examples (terms) as an initial classifier by a user is an important step because of the need for personalization applications. The requirements and biases of a user are represented in the classifier.

For example, we have a text database in which there are a lot of mixed texts on soccer teams, software marketing, hot-spring, etc. And this database has been pre-processed by using morphological analysis. Thus we may use the text classification method stated above to obtain the classified sub-databases on soccer teams, software marketing, hot-spring, respectively, as shown in Figure 1.

3 Generation of ontology

Based on the result of text classification, the process of generation of ontology can be divided into the following two major stages.

The first stage is *conceptual relationship analysis* [12, 3]. We first compute the combined weights of terms in texts by Eqs. (4) and (5), respectively.

$$D_i = \log d_i \times tf_i \tag{4}$$

$$D_{ij} = \log d_{ij} \times tf_{ij} \tag{5}$$

where d_i and d_{ij} are the text frequency, which represent the numbers of texts in a collection of n texts in which term i occurs, and both term i and term j occur, respectively, tf_i and tf_{ij} are the term frequencies, which represent the numbers of occurrences of term i, and both term i and term j, in a text, respectively.

Then a network-like concept space is generated by using the following equations to compute their similarity relationships.

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

Table 1: The similarity relationships of the terms

Term i	Term j	$Rel(i,j)$
team	soccer	0.7385
league	soccer	0.7326
university	soccer	0.5409
player	soccer	0.4929
Japan	soccer	0.4033
region	soccer	0.4636
game	soccer	0.1903
sports	soccer	0.1803
gymkhana	soccer	0.1786
soccer	team	0.7438
league	team	0.8643
university	team	0.5039
player	team	0.1891
Japan	team	0.1854
region	team	0.1973
...	...	...

$$Rel(i,j) = \frac{D_{ij}}{D_i} \tag{6}$$

$$Rel(j,i) = \frac{D_{ji}}{D_j}. \tag{7}$$

Here Eq. (6) and Eq. (7) compute the relationships from term i to term j, and from term j to term i, respectively. We also use a threshold value to ensure that only the most relevant terms are remained. Table 1 shows a portion of the similarity relationships of the terms on soccer teams.

The second stage is to generate the prototype of the ontology by using a variant of the Hopfield network. Each remaining term is used as a neuron (unit), the similarity relationship between term i and term j is taken as the unidirectional, weighted connection between neurons. At time 0,

$$\mu_i(0) = x_i : 0 \leq i \leq n-1$$

where $\mu_i(t)$ is the output of unit i at time t, and x_i indicates the input pattern with a value between 0 and 1. At time 0, only one term receive the value 1 and all other terms receive 0. We repeat to use the following equation n times (i.e. for n terms).

$$\mu_i(t+1) = f_s[\sum_{i=0}^{n-1} t_{ij}\mu_i(t)], \qquad 0 \leq j \leq n-1 \tag{8}$$

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)

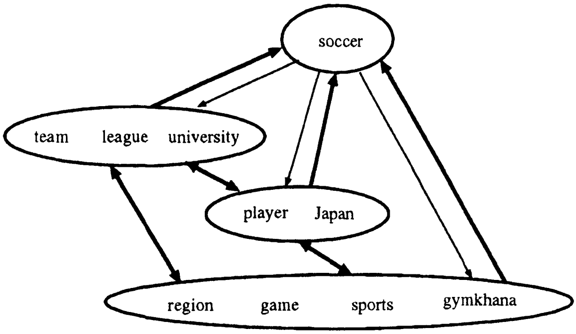

Figure 2: The prototype of a domain specific ontology on soccer teams

where t_{ij} represents the similarity relationship $Rel(i,j)$ as shown in Eq.(6) (or Eq.(7) for t_{ji}), f_s is the sigmoid function as shown below.

$$f_s(net_j) = \frac{1}{1 + exp[(\theta_j - net_j)/\theta_0]} \tag{9}$$

where $net_j = \sum_{i=0}^{n-1} t_{ij}\mu_i(t)$, θ_j serves as a threshold or bias, and θ_0 is used to modify the shape of the sigmoid function.

This process is repeated until there is no change between two iterations in terms of output, that is, it converged by checking the following equation:

$$\sum_{j=0}^{n-1}[\mu_j(t+1) - \mu_j(t)]^2 \leq \varepsilon \tag{10}$$

where ε is the maximal allowable error.

The final output represents the set of terms relevant to the starting term, which can be regarded as the prototype of a domain specific ontology. Figure 2 shows an example of the prototype of a domain specific ontology on soccer teams. It is generated by using each term shown in Table 1 as a starting input pattern for learning on the Hopfield network.

4 Refinement of ontology

There is often a limit to the construction of ontology from text databases, whatever the technique employed. Incorporating any associated knowledge significantly increases the efficiency of the process and the quality of the ontology generated from the text data. A thesaurus is a useful source to be used as a background knowledge base for refinement of ontology. By using the thesaurus, the terms are extended by including their synonym, wider and narrow sense of the terms. Figure 3 shows a domain specific ontology

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

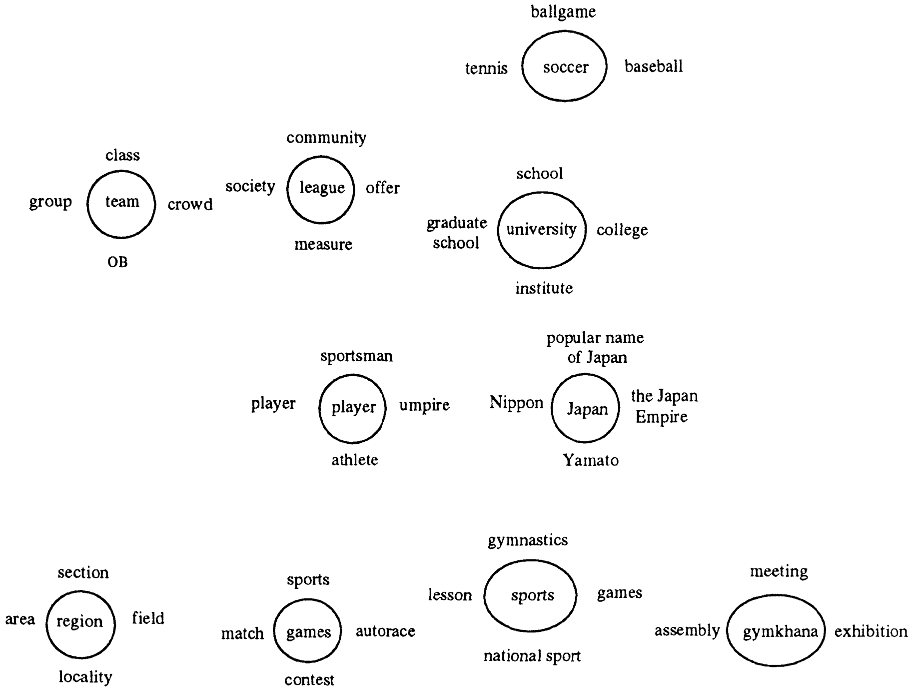

Figure 3: A domain specific ontology refined by using thesaurus

on soccer teams refined by using thesaurus. Note the concept space shown in Figure 3 is the same as the one shown in Figure 2, but we omitted the links between terms for clarity of the figure.

5 Conclusion

The paper presented a multi-phase process of automatic construction of the *domain specific* ontology from text databases, in which text mining and natural-language understanding methods are used. We stress that the process is iterative, and may repeat at different intervals when new/updated data come. Hence how to handle change is an important issue related to refinement of ontology. In particular, during the (long) lifetime of an application session, there may be many kinds of changes such as changes in the text data, the purpose of using both the text data and the ontology, etc. Hence we need a method to reuse the exiting ontology with local adjustment adapted to the changes. In addition, some possible steps such as morphological analysis and generation of classification rules in this multi-phase process were not discussed in this paper although they also are important ones for

Data Mining II, C.A. Brebbia & N.F.F. Ebecken (Editors)
© 2000 WIT Press, www.witpress.com, ISBN 1-85312-821-X

a whole process.

References

[1] Aggarwal, C.C. and Yu, P.S. "On Text Mining Techniques for Personalization", Zhong, N., Skowron, A., and Ohsuga, S. (eds.) *New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, LNAI 1711, Springer-Verlag (1999) 12-18.

[2] Chandrasekaran, B., Josephson, J.R., and Benjamins, V.R. "What Are Ontologies, and Why Do We Need Them?", *IEEE Intelligent Systems*, Vol.14, No.1 (1999).

[3] Chen, H. and Lynch, K.J. "Automatic Construction of Networks of Concepts Characterizing Document Databases", *IEEE Tran. on Sys. Man and Cybernetics*, Vol.22, No.5 (1992).

[4] Chen, H. "Collaborative Systems: Solving the Vocabulary Problem", *IEEE Computer*, Vol. 27, No. 5 (1994).

[5] Cooper, W.S., Gey, F.C., and Dabney. D.P. "Probabilistic Retrieval Based on Staged Logistic Regression", *Proc. ACM SIGIR'92* (1992).

[6] Cooley, R., Mobasher, B., and Srivastavva, J. "Data Preparation for Mining Would Wide Web Browsing Patterns", *Knowledge and Information Systems, An International Journal*, Vol.1, No.1, Springer-Verlag (1999) 5-32.

[7] Frank, G., Farquhar, A., and Fikes, R. "Building a Large Knowledge Base from a Structured Source", *IEEE Intelligent Systems*, Vol.14, No.1 (1999).

[8] Guarino, N. (ed.) *Formal Ontology in Information Systems*, IOS Press (1998).

[9] Ishikawa, Y. and Zhong, N. "On Classification of Very large Text Databases", *Proc. the 11th Annual Conference of JSAI*, (1997) 300-301 (in Japanese).

[10] Lewis, D.D. and Catlett, J. "Heterogeneous Uncertainty Sampling for Supervised Learning", *Proc. Eleventh Inter. Conf. on Machind Learning* (1994).

[11] Nishio, Y., Kakemoto, Y., and Zhong, N. "Multi-Lingual Internet Search Engine with Automatic Ontology Maintenance System", *Proc. PAKDD-99 Workshop on Knowledge Discovery from Advanced Databases (KDAD'99)* (1999).

[12] Salton, G. *Automatic Text Processing*, Addison-Wesley Publishing (1989).

[13] Yao, Y.Y. and Zhong, N. "An Analysis of Quantitative Measures Associated with Rules", Zhong, N. and Zhou, L. (eds.) *Methodologies for Knowledge Discovery and Data Mining*, LNAI 1574, Springer-Verlag (1999) 479-488.

[14] Zhong, N., Kakemoto, Y., and Ohsuga, S. "An Organized Society of Autonomous Knowledge Discovery Agents", Peter Kandzia and Matthias Klusch (eds.) *Cooperative Information Agents*. LNAI 1202, Springer-Verlag (1997) 183-194.

[15] *A Survey on Knowledge Discovery from VLDB*, The Japan Research Institute (1996) (in Japanese).