Acquisition of Web Semantics based on a Multilingual Text-Mining Technique

Chung-Hong Lee[1] and Hsin-Chang Yang[2]

[1] Department of Electrical Engineering,
National Kaohsiung University of Applied Sciences,
415 Chien Kung Road, Kaohsiung 807, Taiwan
leechung@mail.cju.edu.tw
[2] Department of Information Management, Chang Jung University,
Tainan, Taiwan
hcyang@mail.cju.edu.tw

Abstract. This paper describes how semantic web issues such as machine-understandable semantic representation in multiple languages and acquisition of web semantics are tacked by means of a multilingual text-mining approach. This paper presents algorithms that enable multilingual text mining based on neural network technique, namely the self-organizing maps (SOM) for automatically grouping similar multilingual texts (i.e. Chinese and English texts). We treat the World Wide Web as our multilingual corpus and utilize the text-mining model for grouping semantically similar documents. When expose to raw multilingual text, the model clusters words and documents according to their semantic relatedness to form a semantic network.. The model also suggests a novel explanation of multilingual text clustering based on the semantic relatedness. In addition, we propose a framework based on the semantics extracted from the resulting document clusters, along with a XML document technique, to carry out the task of re-categorization of web pages. The web pages are re-categorized based on the semantic relatedness in the corpus. For implementation, we applied the derived multilingual text mining approach in the process of automatic construction of a new text directory, in order to a XML document database that achieves the aims of a semantics-based knowledge categorization.

1 Introduction

The "Semantic Web" is used to denote the next evolution step of the Web, which establishes a layer of machine understandable data. The data is suitable for automated agents, sophisticated search engines and interoperability services, which provide a previously not reachable grade of automation. The ultimate goal of the Semantic Web is to allow machines the sharing and exploitation of knowledge in the Web way, i.e. without central authority, with few basic rules, in a scalable, adaptable, extensible manner.

1.1 Semantics in Web Content

Recent work in computational linguistics suggests that large amounts of semantic information can be extracted automatically from large text corpora on the basis of lexical co-occurrence information. Such semantics has been becoming important and useful for being as an essential representation of content in each web page, particularly with the increasing availability of digital documents in various languages from all around the world. In this work we attempt to develop a novel algorithmic approach for multilingual text mining technique to extract semantic information from the web text corpora. Using a variation of automatic clustering techniques, which apply the *Self-Organizing Maps* (*SOM*), we have conducted several experiments to uncover associated documents based on Chinese-English bilingual parallel corpora, and a hybrid Chinese-English corpus. When exposed to raw multilingual text, the model clusters words and documents according to their semantic relatedness to form a semantic network, as a source of metadata for the construction of the Semantic Web.

2 Multilingual Text-Mining for Creating Metadata

Multilingual text mining is the extraction of implicit, previously unknown, and semantically similar terms and documents from multilingual corpora. The objective is to establish computer programs that automatically cluster conceptually related terms and texts in various languages.

In this work, the primary thrust is to make the text mining techniques fully multilingual, mining open-source texts in different languages. However our goal in these experiments is not to establish the ideal full-functional multilingual information discovery system, as the time and resources required for this task are considerable. Rather, we restrict our attention to bilingual corpora and try to understand the basic requirements for effective multilingual information discovery and the problems that arise from a simple implementation of such a system. This research work is mainly carried out by experimenting with text extraction from parallel corpora, a variation of bilingual corpora. Bilingual corpora can be used in many ways: For multilingual information access, bilingual corpora make it possible to investigate syntactic, semantic and lexical relationships between languages and are also important sources of contrastive evidence of language in usage. Generally speaking, there are two different types of bilingual corpora: parallel and comparable corpora. Parallel text corpora are sets of translation-equivalent texts, in which generally one text is the source text and the other(s) are translations. Comparable text corpora are sets of texts from pairs or multiples of languages which can be contrasted and compared because of their common features. However, unlike parallel (i.e. translation-equivalent) corpora, they always concern a restricted sublanguage. They offer a source of data on natural language lexical equivalents within a given domain [17]. For multilingual text mining, in this work we expect it to be acted as a starting point for exploring the impacts on linguistics issues with the machine learning approach to mining sensible linguistics elements from multilingual text collections. It is believed that, if the corpus is large enough, then statistical or other algorithm-based techniques can be used to produce bilingual term equivalents or associations by comparing which strings co-occur in the same sentences over the whole corpus. Therefore in this work we employed a parallel corpus, the bilingual magazine articles from the *Sinorama* Magazine, to create a set of dual-language training documents for discovering cross-language term associations and document associations.

3 SOM based Text-Mining Approach

3.1 Document Preprocessing

The proposed system focuses on the task of finding associations in collections of text. Based on association, similar documents, through the proposed text mining process, can be gathered in a cluster. For an English corpus, the document preprocessing is quite straightforward. Our approach begins with a standard practice in information retrieval (IR) [19] to encode documents with vectors, in which each component corresponds to a different word, and the value of the component reflects the frequency of word occurrence in the document. In practice, the resulting dimensionality of the space is often tremendously huge, since the number of dimensions is determined by the number of distinct indexed terms in the corpus. As a result, techniques for controlling the dimensionality of the vector space are required. Such a problem could be solved by eliminating some of the most common and some of the rarest words, and by applying a numerical algorithm such as *Latent Semantic Indexing (LSI)* [2] method. For a Chinese corpus, the document preprocessing is relatively complicated. Since a Chinese sentence is composed of characters without boundaries, segmentation is indispensable. We employ a dictionary, some morphological rules and an ambiguity resolution mechanism for segmentation. In addition, we also extract named organizations, people, and locations, along with date/time expressions and monetary and percentage expressions. The rest of the process is the same as that of text mining in an English corpus.

3.2 Self-Organizing Maps

The *self-organizing map* (*SOM*) [9][10] is one of the major unsupervised artificial neural network models. It basically provides a way for cluster analysis by producing a mapping of high dimensional input vectors onto a two-dimensional output space while preserving topological relations as faithfully as possible. After appropriate training iterations, the similar input items are grouped spatially close to one another. As such, the resulting map is capable of performing the clustering task in a completely unsupervised fashion. In this work we employ the SOM method to produce two maps for text mining, namely the *word cluster map* and the *document cluster map*.

3.3 The Word Cluster Map and Document Cluster Map

The word cluster map that is employed for document encoding is produced according to word similarities, measured by the similarity of the co-occurrence of the words. Conceptually related words tend to fall into the same or neighboring map nodes. By means of the SOM algorithm, word clusters can be ordered and organized as nodes on the map. Let $\mathbf{x}_i \in \Re^N$, $1 \le i \le M$, be the feature vector of the i^{th} document in the corpus, where N is the number of indexed terms and M is the number of documents. We used these vectors as the training inputs to the map. The map consists of a regular grid of processing units called *neurons*. Each neuron in the map has N synapses. Let $\mathbf{w}_j = \{w_{jn} | 1 \le n \le N\}$, $1 \le j \le J$, be the synaptic weight vector of the j^{th} neuron in the map, where J is the number of neurons on the map. We trained the map by the SOM algorithm:

Step 1. Randomly select a training vector $\mathbf{x}_i$ from the corpus.

Step 2. Find the neuron j with synaptic weights $\mathbf{w}_j$ which is closest to $\mathbf{x}_i$, i.e.

$$\left\| \mathbf{x}_i - \mathbf{w}_j \right\| = \min_k \left\| \mathbf{x}_i - \mathbf{w}_k \right\|. \tag{1}$$

Step 3. For every neuron l in the neighbor of node j, update its synaptic weights by

$$\mathbf{w}_l^{\text{new}} = \mathbf{w}_l^{\text{old}} + \alpha(t)(\mathbf{x}_i - \mathbf{w}_l^{\text{old}}), \tag{2}$$

where $\alpha(t)$ is the training gain at time stamp t.

Step 4. Increase time stamp t. If t reaches the preset maximum training time T, halt the training process; otherwise decrease $\alpha(t)$ and the neighborhood size, and go to Step 1.

The training process stops after time T which is sufficiently large that every feature vector may be selected as training input for certain times. The training gain and neighborhood size both decrease when t increases.

After the training process, the map forms a Word Cluster Map by labeling each neuron with certain words. For the n^{th} word in the corpus we construct an N-dimensional vector $\mathbf{v}_n$ in which only the n^{th} element is non-zero. To label the neurons, we present each $\mathbf{v}_n$ to the map and find the best matching neuron. Since the number of neurons is generally much smaller than the number of words, each neuron in the map may have multiple labels. We may say that a neuron forms a word cluster because the closely related words will map to the same neuron. The word cluster map autonomously clusters words according to their similarity of co-occurrence. Words that tend to be found in the same document will be mapped to close neurons in the map. For example, the Chinese words for "neural" and "network" often occur simultaneously in a document. They will map to the same neuron, or neighboring neurons, on the map. Words that do not occur in the same document will map to distant neurons on the map. Accordingly we can define the relationship between two words according to their corresponding neurons in the word cluster map, and the mining task will be performed based on such relationships. The trained map also forms a Document Cluster Map by labeling each neuron with certain documents. The document feature vectors $\mathbf{x}_i$ are presented to the map to label the neurons. Documents with similar keywords will map to the same or neighboring neurons. The similarity between two documents may be calculated by measuring the Euclidean distance between their mapped neurons in the map. Since the number of the neurons is much less than the number of the documents in the corpus, multiple documents may map to the same neuron. Thus a neuron forms a document cluster. Besides, neighboring neurons represent document clusters of similar meaning, i.e. high keyword co-occurrence frequency.

3.4 Discovery Algorithm

In this section we explain how the word cluster map and document cluster map effectively model the relationship between the words and documents. We transform a document to a vector of word occurrence. After the self-organizing process, two documents will map to near neurons if they contain similar word occurrences. When different words are labeled on the same neuron or near neurons on the word cluster map, they tend to occur in a restricted set of documents. On the other hand, if two words seldom co-occur in any document, they should not be labeled on near neurons. This is because the neuron may be viewed as representing a virtual document containing those words labeled on it. Two words will be mapped to the same neuron if, and only if, they often co-occur in the same document, otherwise the virtual document may not contain these words simultaneously. Neighboring neurons in the word cluster map represent word clusters containing similar words, i.e. words tend to co-occur in the same document. Hence the self-organizing map may measure the word co-occurrence similarity among documents. We define the similarity between the *p*th word and the *q*th word as follows:

$$D_1(p,q) = (1 + 2^{\left\| G(N_p) - G(N_q) \right\|})^{-1} \tag{3}$$

where *Np* is the neuron labeled by the *p*th word and *G*(*Np*) is the two-dimensional grid location of *Np*. Such similarity measures the likelihood of the co-occurrence of words. Large similarity reveals that the two words often co-occur in the same set of documents, which may be considered as a kind of association pattern among the words.

On the document cluster map a neuron represents a document cluster that contains documents of similar meaning, which is defined by the set of highly co-occurring words they contain. Since we train the map using the encoded document feature vectors as input, the weight vector of a neuron represents the occurrence of the words in a virtual document. Such a virtual document may be used as the centroid of the document cluster associated with that neuron. On the document cluster map only those documents containing overlapping words may map to the same neuron. Documents containing non-overlapping words may map to distant neurons. Neighboring neurons represent document clusters with similar (overlapping) sets of words; thus the co-occurrence of words may be determined by the neighborhood spatially. For any document, we can find its similar documents by examining its mapped neuron and neighboring neurons in the document cluster map. The similarity between the *i*th and *k*th document is defined as follows:

$$D_2(i,k) = (1 + 2^{\|G(N_i) - G(N_k)\|})^{-1} \tag{4}$$

where *Ni* here is the neuron labeled by the *i*th document and *G*(*Ni*) is grid location of *Ni* as in Eq.3.

Documents which contain the same sets of words will definitely map to the same neuron, resulting in high similarity defined in Eq.4. Moreover, even if these documents do not contain exactly the same set of words, we may still say that they are conceptually similar because 1) they still contain common words that often co-occur in these documents, and 2) the dissimilar words are likely occurring in documents mapped to the nearby neurons. Document cluster maps provide an effectively way to form document clusters. A neuron on the map represents a document cluster. We can also define the similarity between two clusters by the distance of their corresponding neurons:

$$D_3(j,l) = (1 + 2^{\|G(N_j) - G(N_l)\|})^{-1} \tag{5}$$

where *j* and *l* are the neuron indices of the two clusters.

The similarities defined in Eq.3, 4, and 5 provide some knowledge about the documents in the corpus. We use Eq.3 to discriminate words based on the knowledge discovered from the corpus. This is also true for Eq.4 and Eq.5 to discriminate documents and clusters respectively. Word associations, as well as document associations, are clearly defined by such similarities. It is natural to apply such associations to applications of document retrieval, indexing, and clustering. A query, which is formulated as a keyword or a combination of keywords adjacent by Boolean operators, is sent by the user to retrieve relevant documents. We perform a search through the word cluster map to obtain the labeled neurons of the keyword or combinations of keywords. The documents labeled on the corresponding neurons on the document cluster map are selected and presented to the user. Moreover, all documents may also be presented by document similarity defined in Eq.4. Documents with higher similarities appear earlier in the query result. The user gets a list of relevant documents even when the query words may not occur in the documents. Important information and new knowledge related to the user's query may be revealed by our method.

4 Multilingual Text Mining from a Hybrid Chinese-English Corpus

Furthermore, we tested the developed discovery algorithm against articles of a hybrid Chinese-English corpus, generated by mixing the Chinese web-documents with the English web-documents from the selected bilingual magazine articles in the *Sinorama* Magazine. In the corpus, each bilingual document pair is split into two documents: an English and a Chinese document. In this case, each training document in either Chinese or English in the corpus is deconstructed into a bag of Chinese and English terms respectively. Although the document pairs contained in the parallel corpus are translation-equivalent, we regard each one of the document pair as an independent training document. The developed SOM method treats this document as a bag of freely intermingled Chinese and English terms. The Chinese-text part of the corpus was first preprocessed by the word segmentation program as in previous experiments. For English documents, the document preprocessing still begins with term indexing, establishing stop lists, stemming, and then encoding documents with binary vectors, in which each component corresponds to a different word, and the value of the component reflects the presence of word in the document.

We constructed self-organizing maps that contain 36 neurons in 6×6 grid format to conduct experiments based on a small collection of the corpora (58 Chinese articles and 58 English articles). The resulting maps give the impression that the effectiveness of text clustering is quite significant in a mixture of English and Chinese experimental articles. An example of resulting word clusters from the trained word cluster map is shown in Table 1.

sinorama 工作 光華 bad bridg caught childlik chingju chrissi commun comprehens contribut countryw cultiv curios drove easili endlessli event eventu fulfil goal greatest highest impart inexhaust inferior jiafong joi magazin mission model modesti plant potenti problem profession pursu record repres respons scholar sens serv spark specialist transmit wai wang wonder 人中 人生 不只 不亞於 不斷 之間 充當 本國 生態 目的 丟人 她們 成就 自然 似乎 我們 赤忱 使命感 其他 委員 孩子 後來 後進 既然 根基 留下 追求 做好 執著 培養 專家 帶動 啓發 深厚 現在 責任 這些 提到 傳遞 敬業樂群 楷模 態度 榮譽 撰述 潛力 稿子 學者 擔任 樹立 橋樑 環保 總是 總編輯 謙遜 職守 灌輸

Table 1. An example of resulting word clusters from the trained word cluster map.

5 System Framework

The system framework is shown in Fig. 1. By applying the SOM neural-network technique, we extract semantics from the contents of a huge HTML-document collection using the developed multilingual text-mining algorithm [12][13]. The extracted semantics can be used to produce metadata and ontology to describe the information in the original web documents. Therefore, the original web documents can be transformed into XML documents and stored in the XML document database with the developed ontology for text categorization. As such, the existing web pages can be analyzed to generate a set of metadata to describe their content and produce an ontology for the XML document database through a text-mining technique, based on their semantic relatedness.

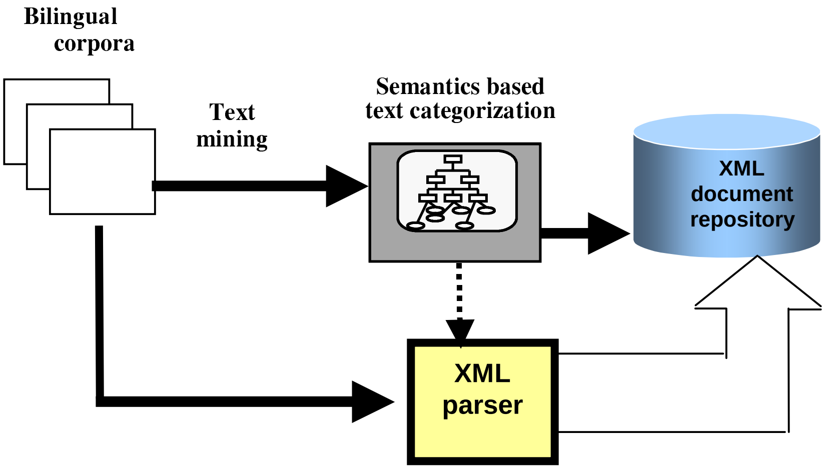

Fig. 1. System Framework

6 Related Work

Text mining is a new interdisciplinary field. It combines the disciplines of data mining, information extraction, information retrieval, text categorization, machine learning, and computational linguistics to discover structure, patterns, and knowledge in large textual corpora. With the huge amount of information available online, the World Wide Web has been becoming a fertile area for text mining research. The Web content data include unstructured data such as free texts, semi-structured data such as HTML documents, and a more structured data such as tabular data in the databases. However, much of the Web content data is unstructured text data. The research around applying knowledge discovery techniques to unstructured text is termed knowledge discovery in texts (KDT) [3], or text data mining [6], or text mining. Advances in computational resources and new statistical algorithms for text analysis have helped text mining develop as a field. Recently there have been some innovative techniques developed for text mining. For example, Feldman uses text category labels (associated with Reuters newswire) to find unexpected patterns among text articles [1][3][4][5]. Text mining by using *self-organizing map* (*SOM*) techniques has already gained some attentions in the knowledge discovery research and information retrieval field. The paper of [16] perhaps marks the first attempt to utilize SOM (unsupervised neural networks) for an information retrieval work. In this paper, however, the document representation is made from 25 manually selected indexed terms and is thus not really realistic. In addition, among the most influential work we certainly have to mention WEBSOM [7][8][11]. Their work aims at constructing methods for exploring full-text document collections, the WEBSOM started from Honkela's suggestion of using the *self-organizing semantic maps* [18] as a preprocessing stage for encoding documents. Such maps are, in turn, used to automatically organize (i.e., cluster) documents according to the words that they contain. When the documents are organized, following the steps in the preprocessing stage, on a map in such a way that nearby locations contain similar documents, exploration of the collection is facilitated by the intuitive neighborhood relations. Thus, users can easily navigate a word category map and zoom in on groups of documents related to a specific group of words. Differing from above traditional utilization of Self-Organizing Maps in text mining, in this project (including previous work [12][14][15][20][21]) we did not employ the map representation for visualization. Rather, we provide a concept-based query method for document retrieval from the database, in order to effectively improve the difficulties in navigating the maps for text search. Our system allows user to perform a keyword-based search similar to that of other search engines. Once user input a keyword, the input term was used as an entry point into the generated word cluster and the searching tool would return a list of suggested documents that were related to the input term, according to the SOM clustering algorithm. With particular emphases on processing multilingual information sources, the developed SOM technique performs a language-neutral method to tackle the linguistics difficulties in the text mining process. The algorithm was presented in detail in our previous work [12][14][15][20][21].

7 Conclusions

Still, in this case text mining from the "hybrid Chinese-English corpus" text collection actually represents typical similarity rather than translation equivalence. However, in this case terms including both Chinese and English ones which are conceptually similar are grouped in a neuron and neighboring neurons. In addition, an interesting outcome is that the resulting cluster groups words with similar concepts which are represented either in English or Chinese. This implies the developed mining algorithm is neutral to languages (i.e. English and Chinese) used in the experimenting documents. This potential provides possibilities to establish an effective way for concept extraction from multilingual information sources. Furthermore, the resulting term clusters can be used to establish a structure of concepts or an ontology. The ontology is often language-independent, or at least language-neutral, so that it can be used in multilingual Semantic Web applications.

Acknowledgement. This research has been partially supported by the National Science Council in Taiwan (NSC 90-2213-E-309-001).

References

1. Dagan, I., Feldman, R. and Hirsh, H."Keyword-Based Browsing and Analysis of Large Document Sets". Proc. of the Fifth Annual Symposium on Document Analysis and Information Retrieval (SDAIR), Las Vegas, NV, (1996).
2. Deerwester, S., Dumais, S., Furnas, G. and Landauer, K. "Indexing by Latent Semantic Analysis," Journal of American Society for Information Science 40(6), (1990) 391-407.

3. Feldman, R. and Dagan, I. “KDT - Knowledge Discovery in Texts”. Proc. of the First Annual Conference on Knowledge Discovery and Data Mining (KDD), Montreal (1995).
4. Feldman, R., Klosgen, W. and Zilberstein, A. “Visualization Techniques to Explore Data Mining Results for Document Collections”. Proc. of the Third Annual Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach (1997).
5. Feldman, R., Dagan, I. and Hirsh, H. “Mining Text Using Keyword Distributions”. J. of Intelligent Information Systems, Vol. 10, (1998) 281-300.
6. Hearst, M.A. “Untangling Text Data Mining”. Proceedings of ACL’99: the 37th Annual Meeting of Association for Computational Linguistics, (1999).
7. Honkela, T., Kaski, S., Lagus, K. and Kohonen, T. “Newsgroup Exploration with WEBSOM Method and Browsing Interface”. Technical Report A32. Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland (1996).
8. Kaski, S., Honkela, T., Lagus, K. and Kohonen, T. “WEBSOM--Self-Organizing Maps of Document Collections. Neurocomputing”, Vol. 21 (1998) 101-117.
9. Kohonen, T. “Self-Organizing Formation of Topologically Correct Feature Maps,” Biological Cybernetics, Vol. 43, pp.59-69, (1982).
10. Kohonen, T. Self-Organizing Maps, Springer-Verlag, Berlin, 1995.
11. Kohonen, T. “Self-Organization of Very Large Document Collections: State of the Art”. In Niklasson, L., Boden, M., and Ziemke, T., editors, Proc. of ICANN98, the 8th International Conference on Artificial Neural Networks, Vol. 1, London, (1998) 65-74. Springer.
12. Lee, C.H. and Yang, H.C., “Towards Multilingual Information Discovery through a SOM based Text Mining Approach”. In Proceedings of International Workshop on Text and Web Mining, The Sixth Pacific Rim International Conference on Artificial Intelligence (PRICAI 2000), Melbourne, Australia, August 28-September 1, (2000) 81-87.
13. Lee, C.H. and Yang, H.C., “A Multilingual Text Mining Approach Based on Self-Organizing Maps”. In *Applied Intelligence* (The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies): Special Issue on Text and Web Mining. (SCI)(EI). In Press. (2003).
14. Lee, C.H. and Yang, H.C., “A Web Text Mining Approach Based on Self-Organizing Map”. Proc. of the ACM CIKM'99 2nd Workshop on Web Information and Data Management (WIDM'99), Kansas City, Missouri, USA, (1999) 59-62.
15. Lee, C.H. and Yang, H.C., “A Text Data Mining Approach Using a Chinese Corpus Based on Self-Organizing Map”. Proc. of the Fourth International Workshop on Information Retrieval with Asian Language (IRAL'99), Taipei, Taiwan, (1999) 19-22.
16. Lin, X., Soergel, D. and Marchionini, G. “A Self-Organizing Semantic Map for Information Retrieval”. Proc. of the ACM SIGIR Int’l Conf on Research and Development in Information Retrieval (SIGIR’91), Chicago, IL (1991).
17. Picchi, E. and Peters, C. “Cross-Language Information Retrieval: A System for Comparable Corpus Querying”. In Cross-Language Information Retrieval. eds. by Grefenstette, G. Kluwer Academic Publishers, (1998) 81-92.
18. Ritter, H. and Kohonen, T. “Self-Organizing Semantic Maps”. Biological Cybernetics, Vol. 61 (1989) 241-254.
19. Salton, G. and McGill, M.J. Introduction to Modern Information Retrieval, McGraw-Hill Book Company, New York, 1983.
20. Yang, H.C. and Lee, C.H. (2000), “Automatic Category Generation for Text Documents by Self-organizing Maps”. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN 2000), Como, Italy, July 24-27, 2000. Vol. III-581-586.
21. Yang, H.C. and Lee, C.H. (2000), “Automatic Category Structure Generation and Categorization of Chinese Text Documents”. In The Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD-2000), Lyon, France, Sep. 13-16, 2000.