Concept Extraction and Clustering for Topic Digital Library Construction

Zhang Chengzhi[1, 2], Wu Dan[3]
1. Department of Information Management, Nanjing University of Science & Technology, Nanjing 210094;
2. Institute of Scientific & Technical Information of China, Beijing 100038;
3. School of Information Management, Wuhan University, Wuhan 430072.
zhangchz@istic.ac.cn, woodan@whu.edu.cn

Abstract

This paper is to introduce a new approach to build topic digital library using concept extraction and document clustering. Firstly, documents in a special domain are automatically produced by document classification approach. Then, the keywords of each document are extracted using the machine learning approach. The keywords are used to cluster the documents subset. The clustered result is the taxonomy of the subset. Lastly, the taxonomy is modified to the hierarchical structure for user navigation by manual adjustments. The topic digital library is constructed after combining the full-text retrieval and hierarchical navigation function.

1. Introduction

The organization methods of information play an important role in the application service of the Internet. Under the Internet environment with massive data, the traditional methods cannot answer users' information needs adequately and timely. At the same time, the artificial intelligence techniques have irreplaceable function in the application service of Internet. However, it is difficult to response the service request in time due to the high-dimensional data computation. Meanwhile, because of lacking the mechanism of semantic understanding, there are a lot of information noises..

To resolve these difficulties, it is urgent to integrate the information organization methods with the learning methods of artificial intelligence techniques. Document clustering based on concept or subject method emerges as the times require through the integration of the subject method and the clustering analysis method. Concept extraction is one of basic tasks in the information extraction, and document clustering based on concept is the process of information clustering using the result of concept extraction.. Topic digital library (TDL) is an important application service and it is a special domain digital library based on concept or subject features. This paper will discuss the design and implementation of a TDL system based on concept extraction and document clustering.

2. Related Works

Some works related to TDL construction include SOMLib [1], Scatter/Gather [2] etc..

For topic digital library construction, we can divide the construction process into three sections as follows.

(1) Concept Extraction: Existing methods about concept extraction can be divided into three categories, i.e. simple statistics, linguistics, sophisticated statistics. The simple statistics methods include word frequency, TF*IDF [3]. Linguistics approaches use the linguistics feature of the words, sentences and documents, and this approach includes the lexical analysis, syntactic analysis, discourse analysis [4]. Sophisticated statistics methods include the C-value /NC-value method [5].

(2) Concept Clustering: There are some works related to concept clustering. Concept Clustering Knowledge Graphs contain multiple concepts interrelated through multiple semantic relations together forming a semantic cluster represented by a conceptual graph [6]. Kang, Chang & Hsu use keyword to automatic cluster document [7] [8]. Topic-driven Clustering method was proposed by Zhao & George [9].

(3) Clustering Description: Document clustering description is a problem of labeling the clustered results of documents clustering. It can help users determine whether one of the clusters is relevant to users' requests. Existing methods of labeling document clusters include: simple statistics-based method, e.g. TF [2], linguistics resource-based method, e.g. WordNet [10], and other approaches, e.g. DCF [11].

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.81

3. Framework

Concept extraction and document clustering (CEDC) can be divided into three sections: concept extraction, document clustering based on concept and clustering description. The process of CEC includes 6 steps: 1) pre-treatment for clustering objects using lexical analysis, syntactic analysis. 2) concept extraction from clustering objects using extraction model and extraction performance evaluation. 3) concept space generation through text representation model. 4) object similarity computation according to similarity model. 5) object clustering using clustering model and clustering performance evaluation. 6) clustering result description through clustering description model and description performance evaluation.

4. TDL Construction Based on CEDC

This section gives the design of topic digital library and describes the three key technologies in detail, i.e. concept extraction, document clustering based on concept and clustering description.

4.1. Design of Topic Digital Library

The TDL provides information services including information collection, storage, clustering navigation and full-text retrieval for users in the special domain. TDL is designed as follow.

Firstly, documents subset of a special domain is produced by automatic document classification approach. It combines the rule-based and statistical method to classify the documents from the large-scale document collection. Then, the keywords of each document are extracted through the machine learning. The keywords are used to cluster the documents subset. The clustered result is the taxonomy of the subset. Lastly, the taxonomy is modified to the hierarchical structure for user navigation by manual adjustments. The TDL is constructed after combining the full-text retrieval and hierarchical navigation function.

4.2. Key Technologies of Topic Digital Library

As mentioned above, key issues in the process of TDL include automatic document classification, concept extraction, document clustering, data integration etc. Because the technology of automatic document classification is researched widely, the detail about it is not described in this paper. Three key technologies of TDL are detailed as follows.

4.2.1. Concept Extraction. Many automatic concept extraction approaches on a small-scale corpus had been proposed, but few approach involved in massive data sets. This paper combines the simple statistics method and the linguistics feature of the words to extract the concept of the document of massive data sets. We construct a large-scale keyword dictionary using the journal database resources of *CNKI*. The term frequency and inverse document frequency (*TF×IDF(t)*), frequency (*KeyFreq(t))*, diameter (*Diameter(t)*), length (*Length(t))*, position of the first occurrence (*FirstLoc(t)*), distribution deviation (*Deviation(t)*) of the keyword (*t*) inside the document (*D*) is combined to compute the total score (*Weight(t)*) as follows.

$$Weight(t) = TF \times IDF \times KeyFreq(t) \times Diameter(t) \times FirstLoc(t) \times Length(t) \times Deviation(t) \quad (1)$$

Given the number K, we select the top K keywords with the highest scores in a Chinese document as the concepts of the document.

4.2.2. Document Clustering Based on Concept. After concept extraction, the documents set can be represented by concept matrix in the concept space. We use sample weighting clustering algorithm based on K-Means algorithm to group the documents. The algorithm uses academic documents as the clustering objects. In the process of document clustering based on concept, the document and the center of the cluster are represented by the concept matrixes. The similarity between the clustering objects is calculated by the cosine of the angle between the concept matrixes.

In sample weighting clustering algorithm, after weighting the clustering samples, the clustering criterion function is given as follows.

$$\boldsymbol{J'} = \sum_{i=1}^{K}\sum_{j=1}^{m_i}(w_j \cdot Sim(\vec{d}_j, \vec{c}_i')) \quad (2)$$

Where w_j denotes the weight of sample j with the constraint of $\sum_{j=1}^{m_i} w_j = 1$. $\vec{c}_i'$ is the prototype of cluster i after clustering samples are weighted , and it can be computed according to the formula (3).

$$\vec{c}_i' = \sum_{j=1}^{m_i}(w_j \cdot \vec{d}_j) \quad (3)$$

The weight value of each document is calculated according to the cited relationship among them.

4.2.3. Clustering Description. Document clustering description is a problem of labeling the clustered results of document collection clustering. It can help users determine whether one of the clusters is relevant to users' information requests. To resolve the problem

of the weak readability of the traditional documents clustering results, we propose a method of automatic labeling documents clusters based on machine learning.

This paper uses Support Vector Machine model to automatic label the results of document clustering. Because the cluster center is concept matrix, the keyword in the cluster center is considered as candidate clustering description of the current cluster. The features in the process of clustering description include the document frequency and inverse cluster frequency, average value of position of the first occurrence in the current cluster, Part-of-speech, length of keyword.

The hierarchical structure is generated after clustering documents in each cluster. The clustered result is the taxonomy of documents set, and it is modified to the hierarchical structure for user navigation after manual adjustments. .

4.3 Implement of Topic Digital Library

The TDL is designed and implemented based on documents database according to framework of TDL. The topic database is generated after topic collection, concept extraction and document clustering, clustering description. The taxonomy is modified to the hierarchical structure for user navigation by manual adjustments. The topic digital library is constructed after combining the full-text retrieval and hierarchical navigation function. When users query of browser of TDL system, they can use the function of clustering navigation and retrieval the sub-topic of the current topic database. We have developed 10 topic database. The on-line version of TDL system is open and can be found at '*http://topic.cnki.net*'.

5. Evaluation of Topic Digital Library

We try to evaluate the performance of TDL system according to the performance of clustering navigation results, namely, evaluate the hierarchical structure of TDL. It is worth noting that clustering navigation evaluation combines hierarchical structure evaluation in the macroscopic view and clustering description evaluation in the microscopic view.

We designed an '*Evaluation Question Fields*' (*EQF*) to evaluate the performance of the TDL system. The EQF includes three questions as shown in table 1.

Five volunteers were recruited to evaluate the clustering description and score manually according to the EQF of clustering description. Table 1 shows the rule for the scoring. The equilibrium degree of clustering description denotes equilibrium degree of clustering objects distribute in each cluster. Relevance degree of clustering description denotes the relevance degree between the clustering description and the topic of the current TDL. Overall effect of clustering description means the overall evaluation of the volunteers for clustering description.

Table 1. Evaluation Question Fields of Clustering Description

No.	Evaluation Standard	Rule for Score manually
1	Equilibrium Degree	Good(2points), General(1point), Bad(0point)
2	Relevance Degree	Good(2points), Genera(1point), Bad(0point)
3	Overall Effect	Good(7~10 points), General(4~7 points), Bad(0~4 points)

In order to further investigate the performance of clustering description, a baseline method is proposed and evaluated in this paper. The idea of the baseline method (denoted as *BL*) is as follows.

The frequency of keywords in the documents set is computed first. Then, the Top N keywords with the highest frequency are selected as the clustering description of the first level in the hierarchical structure. The frequency of keywords in each cluster is computed and the top M keywords with the highest frequency are selected as the clustering description of the second level in the hierarchical structure.

Table 2. Evaluation Result of Clustering Description

Standard / Domain	Equilibrium Degree		Relevance Degree		Overall Effect	
	CEDC	**BL**	**CEDC**	**BL**	**CEDC**	**BL**
Realty	1.57	1.12	1.78/1.82	1.48/1.50	8.12	6.92
Coal	1.68	1.21	1.72/1.78	1.68/1.71	8.20	7.04
Football	1.45	1.01	1.62/1.67	1.59/1.62	7.38	5.94
Aerospace	1.64	0.92	1.53/1.61	1.37/1.45	7.82	5.46
Automobile	1.49	0.94	1.61/1.70	1.52/1.58	7.49	5.59
Average	**1.57**	1.04	**1.65/1.72**	1.53/1.57	**7.80**	6.19

The subjects evaluated the clustering description of five TDL according to evaluation standard and scoring rules in the table 1. Table 2 shows the evaluation results of clustering description. Where, TC, BL denotes CEDC and baseline method respectively. As shown in the table 2, the relevance degree evaluation is divided into the first level and the second level evaluation in the hierarchical structure. For example, the evaluation result of TDL in the realty domain is '*1.78/1.82*'. Where, the first level relevance in the hierarchical structure is '*1.78*' and the second level relevance in the hierarchical structure is '*1.82*'.

As shown in Table 2, the equilibrium degree of the CEDC is higher than the baseline method. The equilibrium degree of the CEDC is '*1.57*' and the latter

is '*1.04*'. It shows that equilibrium degree of the CEDC and baseline method is both high. Because the latter method can't resolve the problem of cluster overlap, the equilibrium degree of it is lower than CEDC.. The relevance evaluation result of the CEDC is: the relevance of the first level in the hierarchical structure is '*1.65*' and the second level is '*1.72*'. This result is better than the result of the baseline. The CEDC uses multiple features of the candidate clustering description. In the process of CEDC, the clustering description can be selected using the SVM model. The CEDC is better than the baseline method in the standard of overall effect. The score of CEDC is '*7.80*' (7~10points) and the score of the baseline is '*6.19*' (4~7points). It shows that the overall performance of the CEDC is better than the baseline method.

Above all, in the view of clustering description, the CEDC is better than the baseline according to the equilibrium degree, relevance degree and overall effect.

6. Conclusion and Future Work

Topic digital library is a special domain digital library based on topic or concept features. A method to build topic digital library based on concept extraction and document clustering is proposed in this paper.

The future wok includes finding the global optimization in the process of building the topic digital library, investing the evaluation method of the topic digital library in the application service.

Acknowledgements

The work has been supported in part by supported by National Key Project of Scientific and Technical Supporting Programs (NO. 2006BAH03B02), Youth Research Support Fund (NO. JGQN0701) and Scientific Research Starting Foundation funded by Nanjing University of Science & Technology (NO. AB41123), Project of the Education Ministry's Humanities and Social Science funded by Ministry of Education of China (NO. 06JC870001).

References

[1] A. Rauber and D. Merkl. SOMLib: A Digital Library System Based on Neural Networks. Proceedings of the Fourth ACM conference on Digital Libraries, Berkeley, CA, USA, 1999: 240-241.

[2] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/Gather: A Cluster-Based Approach to Browsing Large Document Collections. Proceedings of the 15th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'92), Copenhagen, Denmark, 1992: 318-329.

[3] B. Ribeiro-Neto and R. Baeza-Yates. Modern Information Retrieval. ACM Press / Addison-Wesley, 1999.

[4] S. F. Dennis. The Design and Testing of a Fully Automatic Indexing-searching System for Documents Consisting of Expository Text. In: G. Schecter eds. Information Retrieval: a Critical Review, Washington D. C.: Thompson Book Company, 1967: 67-94.

[5] K. T. Frantzi, S. Ananiadou, and J. ichi Tsujii. The C-Value/NC-Value Method of Automatic Recognition for Multi-Word Terms. In: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, London, UK, Springer-Verlag. 1998: 585-604.

[6] C. Barriere and F. Popowich. Concept Clustering and Knowledge Integration from a Children's Dictionary. In: Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, 1996: 65-70.

[7] Kang S S. Keyword-based Document Clustering. In: Proceedings of the 6th International Workshop on Information Retrieval with Asian Languages, Sapporo, Japan, 2003: 132-137.

[8] Chang H-C, Hsu C-C. Using Topic Keyword Clusters for Automatic Document Clustering. IEEE Transactions on Information and Systems, 2005, E88-D: 1852-1860.

[9] Zhao Y, Karypis G. Topic-driven Clustering for Document Datasets. In: Proceedings of the Fifth SIAM International Conference on Data Mining, St.Louis, Missouri, 2005: 358-369.

[10] Y. H. Tseng, C. J. Lin, H. H. Chen, Y. H. Lin. Toward Generic Title Generation for Clustered Documents. In: Proceedings of the 3rd Asia Information Retrieval Symposium, Singapore, 2006: 145-157.

[11] W. Dawid. Descriptive Clustering as a Method for Exploring Text Collections. PhD Thesis. Poznan University of Technology, Poznań, Poland, 2006.