Research and Analysis of Improved Extraction Based on Information Processing Technology

Yongqin Wei, Yinjing Guo
College of Information and Electrical Engineer
Shan Dong University Science and Technology
Qingdao 266590, China

Journal of Digital Information Management

ABSTRACT: *This paper analyzes the problem of research topics. The key words are extracted from the academic papers published in the key journals based on information processing technology, and these words have the co-occurrence relationship between clustered. Research topics in the academic field analysis applications, researchers can provide a clear outline; in the information retrieval process, you can help clarify the information needs. We will apply the proposed method to ROCLING seminar papers on the data to extract the important field of computational linguistics research topics. The results show that this method can be applied to the special circumstances of domestic academic field, while taking out the key words in English and Chinese. the word has been said that the field of cluster results can also be an important research topic. These results in a preliminary validation of the method proposed in this paper the feasibility. It can also be found that application of computational linguistics research and practice are closely related, taking out many of the words in the cluster and machine translation, speech processing and information retrieval related to the calculation model in the language, grammar patterns and analysis, broken words and statistical language model type is the calculation is the subject of interest to linguists.*

Categories and Subject Descriptors:
I.2.7 [Natural Language Processing Text analysis]: H.1.2 [User/Machine Systems]; Human information processing

General Terms: Information Extraction, Text Processing

Keywords: Natural language, Analysis, Seminar

Received: 10 November 2011, Revised 27 December 2011, Accepted 7 January 2012

1. Introduction

Information retrieval research is focused on the interface between people and information, recent research trend has focused on the user's background knowledge, the problem in the retrieval process in the cognitive and information skill level In order for an academic field of information dissemination comprehensive understanding of the phenomenon, the so-called "area analysis" by the academic field of important academic activities, such as research, publication, conference participation, and so analysis of research personnel used or generated knowledge organization, structure, cooperation patterns, language and forms of communication, information systems and related standards. The research topic areas of analysis can be said that an analysis of priorities for important research topics can acquire knowledge in the field of organization, help to clarify the information needs, quick access to required information. In addition, by taking a systematic approach to research topics and analysis, academic research can show a full face [1-7], to provide early access for new scholars in the field of a reference, or as a field of academic research to guide development, provides in-depth research has been expanded the scope of academic research.

This paper presents an automated extraction method of research topics, from academic papers published in a collection of selected key words, then words with each other based on the same paper in a meaningful co-occurrence phenomenon, identify each papers may have a research topic, as the analysis of this important research topic area basis. We believe that the paper contains a study of a rich vocabulary message subject [8-9]. In a paper published during the author by paper title, abstract, and the words in this article will examine the issues, methods and results and other topics to convey to the reader, and even references cited papers also contain a number of title words and subject-related information; and readers read the paper, they can determine based on these words and their research interests on the relevant, and this information into personal knowledge construction and the structure. To make an example of this thesis in the paper title, abstract, and this article includes many

"academic", "research topics", "paper" so the words in the hope that readers, you can from these words co-occur with the use [10], understanding the subject of our study is drawn from the academic study of important topics, and interested readers, could be used in research and published on the use of it. Further, in an academic field, you can find some of the attention of the research topics related words appear in many papers. In the field of computational linguistics point of view, we can find such as "Corpus", "analysis", "information retrieval" and so the words appear in many papers, these are the important research topics in the field. And with the research group of words related to the topic will be repeated in many papers. Therefore, if the papers published in academic analysis, selection of topics on behalf of the meaning of words, words between the statistical co-occurrence of these phenomena, the use of this information will often appear together in a group of words clump together into a collection of words formed by collection can be seen as a particular research topic. The subject of a paper in the analysis, they can estimate the terms on behalf of the research topic clustering and relevance of the paper, as to judge whether the thesis of this topic [11]. Therefore, this paper attempts to use natural language processing techniques to analyze the academic papers published, the paper appear to confirm the words, extract contains co-occurrence of words in which the message, then word clustering, as a thematic analysis to identify of information.

We will apply the technology developed by the field of computational linguistics domestic theme analysis. Choose to study computational linguistics as the main reason for this interdisciplinary field of research characteristics, and successfully developed the theory and technology to the academic research and actual system and product development. Research in this area [12-16], researchers mainly from the two disciplines of linguistics and computer science, calculator for scientists, the main research work is to construct a practical computer system to deal with the problem of natural language, for example, the machine translation, font recognition [17], speech recognition, information retrieval and so on. Linguist's work lies in the calculation of the theoretical specification and application, used to explain the phenomenon of natural language understanding model and simulation capabilities. Scientists need to rely on calculators work formed linguist linguistic theory to establish a rational and efficient computer systems; the linguist is to use the calculator scientists developed the theory of computing systems to explore the natural language with the law. In this important research area in addition to the calculator method is applied to the theory of natural language, the most notable studies include the use of corpus of the developed theory and use the language of design and development of practice theory system. Therefore, in this area for academic activities, you can observe two different linguistics and computer science scholars in the disciplines of mutual agitation generated can also be observed from theoretical research to technology development, to the practical application of, analysis of the research topic is a challenging and meaningful research. In addition, the other is our familiarity with this area will help the development of research methods, the preliminary results obtained to make a reasonable interpretation and as a reference for the next phase of improvements [18].

Use ROCLING one to fourteen in the Symposium session data, a total of 235, we extracted a total of 343 key words. Research topics obtained after cluster 34 words on behalf of a collection of important research topics. The results showed that the development of word extraction method can extract the key words in English and Chinese, the word has been said that the field of cluster results can also be an important research topic. Preliminary validation of the method proposed in this paper the feasibility. Results from the study, we also found that application of computational linguistics research and practice are closely related, taking out many of the words in the cluster and machine translation, speech processing and information retrieval related to the calculation model in the language, grammar patterns and analysis, broken words and statistical language model type is the calculation is the subject of interest to linguists.

The remaining sections of this paper is as follows: Section II, first described in a number of related studies and research topics proposed in this paper the concept of analytical methods and reasonable, and based on the use of these concepts to design a series of natural language processing technology for research analysis the method. Then, in Sections III and IV carve out the core technology in this way: word extraction and clustering of topics. Section III, we propose a multilingual environment in the key words extraction method, the data can be obtained from the English papers on behalf of a study on the key words. Have proposed a fourth term clustering methods, the use of co-occurrence relationship between words, the clustering of multiple terms to represent the possible research topics; research described in this section and the degree of correlation between the subject and the paper's calculations. Reported in Section V of this analysis method is applied to the domestic results of the study of computational linguistics. Finally, Section VI is the conclusion.

2. MULTI-lingual environments

In order to extract the field of academic research topic to represent the key words, we first confirm the important papers in the English data, and Chinese multi-word phrases, vocabulary words to enhance the message, then choose a representative to study the subject of these terms, as this stage the results. In academic papers, often in the form of the phrase to express an important research topic, for example in the field of computational linguistics papers can be found such as the English "language model", "machine translation" or Chinese "language model", " Machine Translation "and so on. In addition, the Chinese text, the word with no clear boundaries between words, for natural language processing, you need to first

conduct a word, the text may confirm the word. Therefore, to study the subject analysis, the primary job is to confirm the important from the paper in English and Chinese multi-word phrases. However, academic papers often have many new words appear to represent new concepts, methods and techniques, we cannot advance in all areas included all the possible words to make a very complete dictionary, the word break. And the use of word-law rule-based word segmentation method to deal with both Chinese and English text, difficult to integrate applications. Therefore, this paper uses statistical-based approach, in order to simultaneously solve multiple Chinese words and phrases in the English question.

The method used in this paper is as follows: first use of title, abstract, references, title and other papers of all the textual data in a PAT-tree data structure used to store all data in a string of papers and papers which they are located data. In this paper, the use of statistical information, including all data in the string appear in the total frequency, the string appears in the paper, the average frequency and standard deviation, and then the string after the word complexity. The total frequency of occurrences of the string in the field on behalf of the importance of high frequency indicates that the emergence of a string of papers in the field and often of great significance. String in the event of paper, the average frequency and standard deviation for the occurrence of the string used to indicate the importance of the paper, such as equation (1).

$$R_S \overset{def}{=} m_S + \sigma_S \qquad (1)$$

In type (1), mS and σS representing string S in the paper appear in the frequency and the average standard deviation. When the string S, the average frequency exceeds a certain threshold value, this means that the string is likely to occur many times in many papers, these papers are the key words, should be selected out. Although the string S in the papers or the average frequency is low, but appear multiple times in some papers, these papers are the key words, also needs to be selected out, then the string S will have a larger standard deviation σS. Therefore, we can use the string in the event of paper, the average frequency and standard deviation of the sum of the RS represents a string on the importance of the papers appeared, RS The higher the value of the string for the occurrence of the more important papers.

String around the complexity of the word then you can determine whether it is a complete word or part of other words, the string S, then the word before and after the C1S and C2S, respectively, the complexity of such type (2) and (3) shows

$$C_{1S} \overset{def}{=} -\sum_a \frac{F_{aS}}{F_S} \log\left(\frac{F_{aS}}{F_S}\right) \qquad (2)$$

$$C_{2S} \overset{def}{=} -\sum_b \frac{F_{aS}}{F_S} \log\left(\frac{F_{aS}}{F_S}\right) \qquad (3)$$

Type (2) and (3), a and b represents the data in the string S in the paper before any possible word, and then followed by the word, FS, FaS and FSb are the string S, aS, and the emergence of the total Sb the frequency. The type (2) then the word before the case point of view, if there is more the kind of string S, then the word before, and then each word before the closer the number of occurrences when, C1S the higher the value, on the contrary, when the character string before the first pick only one word, C1S value equal to 0, or a word of the opportunities before then compared with other large very long time, the C1S value close to 0, indicating that the string plus the first pick the word may is a word. The greater complexity of the first word then the more likely represents the string of words, rather than an independent part of other words; followed by the word of the situation is the same reason.

String through the above conditions, re-use of stop words cannot appear in the string end to end rule of thumb, and further filtered to incomplete words. In the past experience, such as prepositions and fixed word stop words often appear in the extracted string end to end, such as “noun” and “noun + of” or “to + verb” and phrase structure. However, stop words in the middle of the string represents a specific phrase, such as “*part of speech*”, so this situation will be retained.

Confirm the data in an important paper in English and Chinese multi-word phrases, the build-off with these words dictionary word processing required. We use long-term priority rules and terms of an overall frequency of all the papers off the data word to be sure that all the data in the papers that appear in words. At this point, we divided the word out, including some in English phrases, words, and some Chinese words. In this paper, the data does not need to make sure all papers may be words, but to extract all possible research topics on behalf of the key words, so we filter the words with the following conditions. First, the data in the Chinese word, mostly some of the prepositions, stop words or not in the previous step, the words forming the word, to be filtered out. Second, an overall frequency and type (1) of the RS value is too small words, but also to filter the grounds as previously described. The remaining data words and their papers appear in the next phase of the case is the object of analysis.

3. Multi-cluster Algorithm

First, we will be the last phase extracted words, the use of multiple clusters, cliques cluster algorithm for word cluster. Selected the smallest degree of correlation in the case, we can get the number of words clustered in these words cluster words, the degree of correlation between each other above the selected minimum extent; and word them with other words The relevance of different cluster in more than one collection. The relevance of the terms used in this thesis, calculated as follows: we first calculate the frequency of each word in a paper title, abstract, and reference titles and other information, the words eigenvalue computation. But a smaller amount of data because the

only access to the above-mentioned paper, the words in which the frequency is not too high in order to make low-frequency words difference not too large, the square root of the frequency of words in each essay information a characteristic value. As a result, each word a set of feature vector, the degree of correlation between the calculation of words they can use the feature vector corresponding to the cosine of the angle between cosine value to estimate. (4) and (5), respectively, the feature vector of word A, and the degree of correlation between the words A and B estimates.

$$\vec{v}_A' \overset{def}{=} [\sqrt{f_{1,A}}, \sqrt{f_{2,A}}, ..., \sqrt{f_{N,A}}] \qquad (3)$$

$$R(A,B) \overset{def}{=} \frac{\vec{v}_A \cdot \vec{v}_B}{\|\vec{v}_A\| \|\vec{v}_B\|} \qquad (4)$$

The results obtained after the cliques algorithm is very strict, only the estimate of the current relationship of words have more than one pair of words a certain threshold may be clustered within a collection. The same or similar concepts in the research topic, however, to represent different words, these words do not necessarily appear in the same data in the paper, the use of words co-occurrence of the phenomenon of relevance estimation method will be very little relevance estimates cannot be cliques algorithm to cluster together these words. In the paper, the following two techniques can solve the above problems.

First of all, we use the LSI (Latent Semantics the Indexing) technology for the above feature vector "words - characteristic matrix M, the singular value decomposition, the matrix M decomposed into three matrices, To, So and Do, makes. Here To and Do the left of M, the matrix formed by the right singular vectors, its size t×r and d×r, t and d, respectively, for the words and the number of features, r, compared with the matrix M the rank So for a size r×r diagonal matrix, the singular values of the diagonal line value of M, and in accordance with diminishing arranged. If we want to obtain a rank k matrix, k<= r, and makes the least squares and M closest, you can take So the diagonal line of the first k singular values, resulting in a size of k×k matrix S, To and Do also take the first k row vectors, the formation of the matrix T and D, respectively. t - × k and d × k. Matrix can be calculated. The use of LSI technology, the retrieval process, when the words estimate, since the estimate the inner product value between the original characteristics of the vector MM 'calculation of two or two words, such as (6) said,

$$MM' \approx \hat{M}\hat{M}' = TSD'(TSD')' = TSD'DS'T' = TSS'T' = TS^2T' \qquad (6)$$

Due to the matrix of the row vector D of each other unit orthogonal and DD '= I and S is the diagonal matrix, S' = S, so. Use of SVD to obtain the characteristics of latent semantic structure, making the original co-occurrence relationship is weak or does not exist, and relevant than the estimate very small two related words; you can get a larger estimate.

Secondly, after the cliques cluster algorithm, the results obtained based on the overlap between the circumstances of their members re-cluster. Assumptions between the two clusters of three or more members of the same, and the remaining members of strong word co-occurrence relations, but also with some paper data, we are about these two words of the cluster collection of union, to generate a new cluster. Shown in Figure 2, A, B, C, D, E and F, six related words, according to their Cco-occurrence relationship between cliques cluster, the Cong integrated.

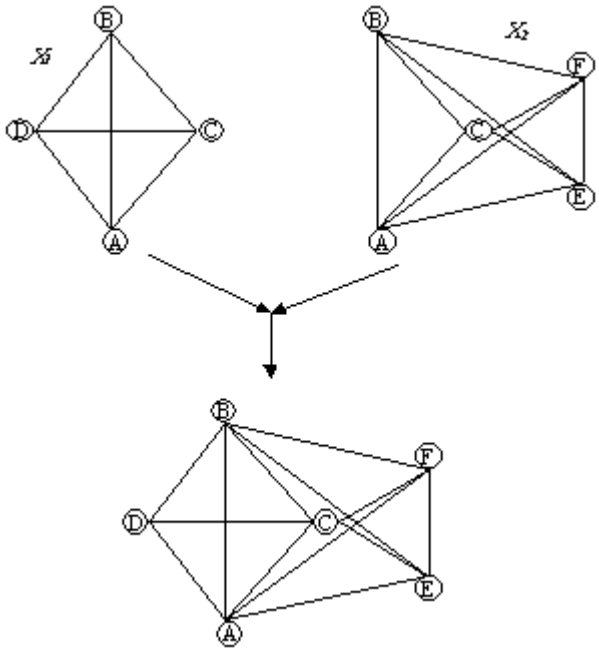

Figure 1. Cluster to Merge two with the Same Members of theC 1. CII. Schematic

Finally, after the cluster processing, you can get some of the important research topics in the representative field of word cluster. Analysis of the research topic, we calculated the correlation between each cluster and paper method of calculation basis for the estimation method for the LSI All papersC, formula (4) to calculate the cluster relevance.

$$R_X = \chi TSD' \qquad (6)$$

X a row vector, each element represents a specific word among the cluster χ other words, if the words included in this cluster, ei the value of 1; otherwise if this cluster does not contain the words ei has the value 0. This is also a row vector of size 1×d, theCresults obtained by the formula (6) R. Xdegree of correlation between each element represents the value of words cluster with the corresponding paper estimates. Finally, according to this result, the relevance of the paper data removed, as research papers relevant to the subject data to analyze.

4. Conclusion

In subsequent studies, in addition to further improve the current method that put forth, in-depth study of topics of the origin, development and evolution will be explored as well as the relevance of the various research themes. Thus

try to test the results graphically manner presented. In addition, the correlation between the different academic areas of research and analysis on the theme is also needed for further study. For example, information retrieval is also of interest in library information science research. Thus how to use natural language processing techniques to analyze the similarities between the two areas is needed.

5. Acknowledgment

This work is a part of the National Natural Science Foundation under Grant. 61071087 and Shandong Province Natural Science Foundation under Grant. ZR2011FM018.

References

[1] Bishop, A. P. Document Structure and Digital Libraries: How Researchers Mobilize Information in Journal Articles, *Information Processing and Management,* 35, p. 255-279.

[2] Lee-Fang Chine, PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval, SIGIR'97, p. 50-58.

[3] Lee-Feng Chien, Chun-Liang Chen, Wen-Hsiang Lu, and Yuan-Lu Chang, Recent Results on Domain-Specific Term Extraction From Online Chinese Text Resources, ROCLING XII, p. 203-218.

[4] KW Church and RL Mercer, Introduction to the Special Issue on Computational Linguistics Using Large Corpora, *Computational Linguistics,* 19 (1), p. 1-24.

[5] LM Covi, Material Mastery: Situating Digital Library Use in University Research Practices, *Information Processing and Management*, 35, p. 293-316.

[6] Deerwester, S., ST Dumais, GW Furnas, TK Landauer, and R. Harshman, Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41 (6) 391-407.

[7]SP Harter, Psychological Relevance and Information Science, *Journal of the American Society for Information Science*, 43 (9) 602-615.

[8] Hatzivassiloglou, L., Gravano, Maganti, A.An Investigation of Linguistic Features and Clustering Algorithms for Topical Document Clustering, SIGIR'2000, p. 224-231.

[9] Hjørland, B., Albrechtsen, H. Towards a New Horizon in Information Science: Domain-Analysis, *Journal of the American Society for Information Science*, 46(6) 400-425.

[10]Chu-Ren Huang, From Quantitative to Qualitative Studies: Developments in Chinese Computational and Corpus Linguistics, Chinese Studies, XVIII special issue, p. 473-509.

[11]GJ Kowalski, MT Maybury, Document and Term Clustering, Information Storage and Retrieval Systems: Theory and Implementation, 2nd ed., Chapter 6, p. 139-163.

[12] Lenders, W. (2012). Past and Future Goals of Computational Linguistics", ROCLING XIV, p. 213-236.
[2] YU J., Zhao Y, Weighted Approximation of Functions with Singularity by q-Baskakov Operators, IEIT Journal of Adaptive & Dynamic Computing, 2012(2), Apr, p. 5-11. DOI=10.5813/www.ieit-web.org/IJADC/2012.2.2

[13] Zhao, C. H., Zhang, J., Zhong, X. Y., Chen, S.J., Liu X. M., (2012)Analysis of Tower Crane Monitoring and Life Prediction, IEIT Journal of Adaptive & Dynamic Computing, 2012(2), Apr, p. 12-16. DOI=10.5813/www.ieit-web.org/IJADC/2012.2.3

[14] Zhao, C. H., Chen, S. J., Liu, X. M., Zhang, J., Zeng, J. (2012). Study on Modeling Methods of Flexible Body in ADAMS, IEIT Journal of Adaptive & Dynamic Computing, 2012(2), Apr, p. 17-22. DOI=10.5813/www.ieit-web.org/IJADC/2012.2.4

[15] Chen, G. Q., Jiang, Z. S., Wu, Y. Q. (2012). A New Approach for Numerical Manifold Method, IEIT Journal of Adaptive & Dynamic Computing, 2012(2), Apr, p. 23-34. DOI=10.5813/www.ieit-web.org/IJADC/2012.2.5

[16] AN Tabah, Information Epidemics and the Growth of Physics, Ph. D. Dissertation of McGill University, Canada.

[14] CL Wayne, Topic Detection and Tracking in English and Chinese, IRAL 5, . 165-172.

[17]TD Wilson, Models in Information Behaviour Research, *Journal of Documentation,* 55 (3) 249-270.

[18] Yang, Y., Pierce, T., Carbonell, J. A Study on Retrospective and On-Line Event Detection, SIGIR'98, p. 28-36.