ISKSS

http://www.jaist.ac.jp/iskss/

Web Text Mining on A Scientific Forum

Wen Zhang[1,2], Xijin Tang[2], and Taketoshi Yoshida[1]

[1] *School of Knowledge Science, Japan Advanced Institute of Science and Technology,*
1-1 Ashahidai, Nomi, Ishikawa 923-1292, Japan
[2] *Institute of Systems Science, Chinese Academy of Sciences, Beijing 100080, P.R. China*

Abstract

XSSC (Xiangshan Science Conference) is well known for its directive role to promote Chinese basic scientific research where the senior experts and authorities of Chinese sciences are invited to attend such conferences and discuss freely concerning the sciences and research in China. In this paper, a Chinese Web text mining process was proposed based on traditional information retrieval and extraction as well as Internet techniques. Then the Web text mining process including texts indexing, summarization and clustering on XSSC Web documents was conducted and the results of these processes were presented with integration using a user interface. Finally, some improvements are indicated.

Keywords: Information extraction, Web text mining, Web text summarization, Web text clustering; XSSC

1. Introduction

The Web offers access to vast amounts of information, but the usefulness of this access is limited by our ability to make sense of Web information in a timely way. With this originality, Web mining is employed to improve our access ability by providing its users with customized, timely and comprehensive information support distilled from Web. Gernerally, Web mining includes Web content mining, Web structure mining and Web usage mining. Web text mining belongs to the area of Web content mining and refers to the process of extracting interesting and non-trivial patterns or knowledge from Web. Usually, it is viewed as an extension of text mining and data mining.

This paper presents a Chinese Web text mining process consisting of four modules: Web crawler, Web content indexer, Web text summarization and Web text clustering. Also, a user interface is proposed to interpret Web text mining results on XSSC. The rest of this paper is organized as follows. Section 2 surveys the prior research on Web text mining. Section 3 gives a Chinese Web text mining process, which is a basic process but includes all the necessary modules to implement a whole Chinese Web text mining. Section 4 describes the practical working details of each module in Web text mining on XSSC and the processing result of each module is presented in this section. The last section concludes this paper and indicates the further research.

2. Prior Research

Recently, a lot of researches in the field of Web text mining have been carried out. On the whole, there are two types research on Web text mining. One is focused on the application of Web text mining techniques to solve practical problems such as patent retrieval, biomatics and the other is focused on the algorithms or the improvements of some Web text mining methods such as Fuzzy method, SVM .etc. For the former aspect, Reference [1] introduced a Web-based news retrieval system AI-Times whose goal is to retrieve and organize the web news information. Reference [2] implemented a medical informatics system Medical Concept Mapper to facilitate access to online medical information sources by providing users with appropriate medical search terms for queries. Reference [3] employed a new concept of seed content to find out the common interest community between individual Web pages. In the latter aspect, Reference [4] brought forward an algorithm for automatic HTML Tags and pure text extraction from semi-structured Web Documents; Reference [5] carried out a knowledge discovery method in texts according to the co-occurrence and distribution of keywords between

texts. While Reference [6] constructed hyperlinks between Web pages automatically using SOM (Self-Organizing Method) training of documents representation vectors.

However, the Web text mining researches mentioned above have their own disadvantages in contrast to their advantages. In their common, they all depend on basic nature language processing (NLP) techniques without any semantic analysis. The application of Web text mining can surely satisfy some practical demands but the success of this kind of application depends on heuristic experience of a certain domain so much that it can not be easily transferred to in other fields because of its lack of feasibility. The research of Web text mining technique have not synthesized both the Web features of texts and the text mining techniques effectively consequently it has been separated into Web representation and text mining instead of the research on the combination of these two aspects.

3. A Chinese Web Text Mining Process

Generally speaking, Web text mining involves two aspects: one is the Web texts acquisition and another is the text mining techniques used on the Web texts. The object of text mining is pure texts while the object of Web text mining is Web pages whose contents contain not only pure texts but also the hyperlinks between Web pages and more preprocess should be counducted on Web texts than pure texts. Figure 1 is a basic Chinese Web text mining process. The function of each module is as follows.

3.1 Web crawler

The role of Web crawler is to collect Web pages from Internet. Here, the traditional spider detecting algorithm listed in the following part is employed with two parameters as the seed URL and the depth for crawling.
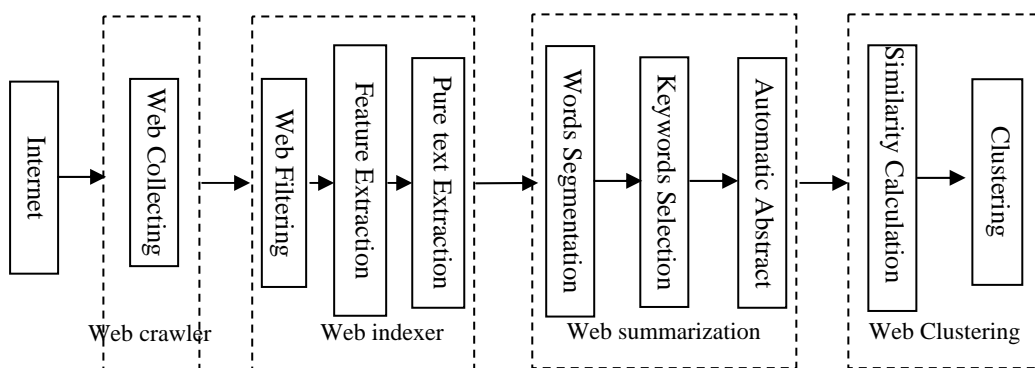

Fig.1. A basic Chinese web text mining process

Firstly, the Web page at the seed URL is downloaded by Web crawler and saved into local file base. Then, all the hyperlinks as URLs are parsed out from the downloaded Web page and saved into local database for a record of the visited URLs. Secondly, the Web pages at the newly parsed URLs are also downloaded into local file base and the URLs as hyperlinks in these newly downloaded Web pages are parsed out for the next step downloading. By this kind of reiteration of downloading and parsing, the Web pages within the reach of the parameter depth from the seed URL are collected. In many cases, restrictions are set for the crawler's downloading by evaluating the URLs parsed out from Web pages so as to derive the Web pages as required. For example, reference [1] classifies the URLs into three types in order to obtain the newly updated news Web pages by examining the parsed URLs at intervals.

```
Algorithm I : Spider Algorithm
Begin
Let I be the a list of initial URLS of the seed website
and D be the depth of the spider will crawl;
Let F be a queue;
        For each URL   i   in I
                Enqueue(i,F);
        End
j=0;
While F is not empty and j < D
        u = Dequeue(F);
        if u has not been processed
Get(u);
        Extract the hyperlinks and relevant caption;
        Let U be the set of hyperlinks extracted;
        For each u in U
             Enqueue(u,F);
        End
    End
```

End
End

3.2 Web content indexer

The information extraction with a predefined format is employed purposely to index the XSSC web documents stored in database and file base. The web page indexer is designed here to filter and reorganize the content of Web pages by extracting features from Web page such as page titles, time stamps, name of person and also the pure texts. In details, it includes three parts: Web filtering, characteristic extraction and pure text extraction. Web filtering is used to exclude the irrelevant Web pages such as advertisement, exception page and so on. Feature extraction is used to extract the related information from Web pages according to the Web structures. Pure text extraction is used to eliminate HTML tags and noises in the Web pages like scripts, cascade style so as to obtain consistent pure text from web page without any HTML tags. By the way, two kinds of methods were currently utilized for extraction. The one is based on the language character type, for example, Chinese and Japanese characters belong to different characters collection in Unicode aggregation. Another is based on the format of Web page. For example, there is a great difference in writing format between HTML language and natural language. Our extraction method in this paper is based on the format of web page.

3.3 Web text summarization

Web text summarization is introduced to extract the representative sentences from the pure text of Web page and the extracted sentences were rearranged as the abstract of the Web page. It is one of the tasks of text mining. In details, Web summarization includes three parts: segmenting the sentences of texts into individual words, keywords selection and automatic abstract. Some languages such as Chinese, Korean and Japanese need to be segmented, other Latin series languages such as English, French and so on do not need to be segmented because one word is an index term in these languages. Keywords selection is used here to select the words from one text which have the most resolving power for distinguishing the text from other texts. Sometimes, keywords selection is closely relevant with the text content and subject, and also it may rely on our subjective experience on words expression. Automatic abstract is used to select and rearrange the representative sentences to construct the abstract of text with coherence and fluency. That is, it constructs a short passage with selected sentences from the original text according to the text compression ratio to express the meaning of original text. In the last few decades, researchers have brought up a number of automatic abstract methods. Basically,

these methods fall into two categories. One is based on the structure and semantics of the source text and another is based on the statistical method such as the frequency of the sentences or words occurrence. In the current state of affairs, it is better to use statistical information in automatic extraction [7].

3.4 Web text clustering

Web text clustering is utilized in this step to cluster the Web pages without supervision according to the similarities between them. Many kinds of motivations are required to cluster the Web pages. Firstly, it can exhibit to its user a whole vision of the information after clustering the Web pages into clusters. Secondly, it can provide user related Web pages from the same cluster automatically when user browses one Web page of this cluster in case of ignoring important information. In order to cluster the Web texts, they should be represented by quantitative vectors so as to be clustered. Mostly, one text could be represented by the vector space model (VSM) or Boolean model [8]. VSM records the frequency of a term in a text whereas Boolean model only record the occurrence or absence of the terms in a text while ignoring the frequency of them. By this kind of representation for texts, a clustering can be done among the all texts.

4. Web Text Mining on XSSC

4.1 System overview of web text mining on XSSC

Based on the Chinese Web text mining process proposed in Section 3, a practical case with Web text mining was conducted and implemented on XSSC Website (http://www.xssc.ac.cn). Figure 2 is the system overview of the Web text mining architecture on XSSC.

4.2 Web crawling on XSSC

We set "http://www.xssc.ac.cn/Web/ListConfs" as seed URL and set the depth as 10 for Web crawler. After the procedure of Web crawler, we obtained 646 Web pages from the XSSC Website.

4.3 Web indexing on XSSC

In this step, those Web pages whose contents include general description of XSSC were eliminated and only the Web pages about the conference details were retained for the later text mining process as their HTML schemas can be easily identified so that the overall information about a certain conference such as conference topic, the utterance records of experts and the conferences participants could be extracted from the web pages in order to construct the indexes for them. After the Web filtering, 208 Web pages whose

contents were within our mining interest were retained in this step.

Figure.3 is a typical Web page from XSSC Website about conference details. By identifying the structure of the Web pages, the information about conference topic, conference content as well as the information about the conference participants can be extracted with a predifined format. Figure.4-6 are the extraction results from the XSSC Web page according to the page structure described in Figure.3.
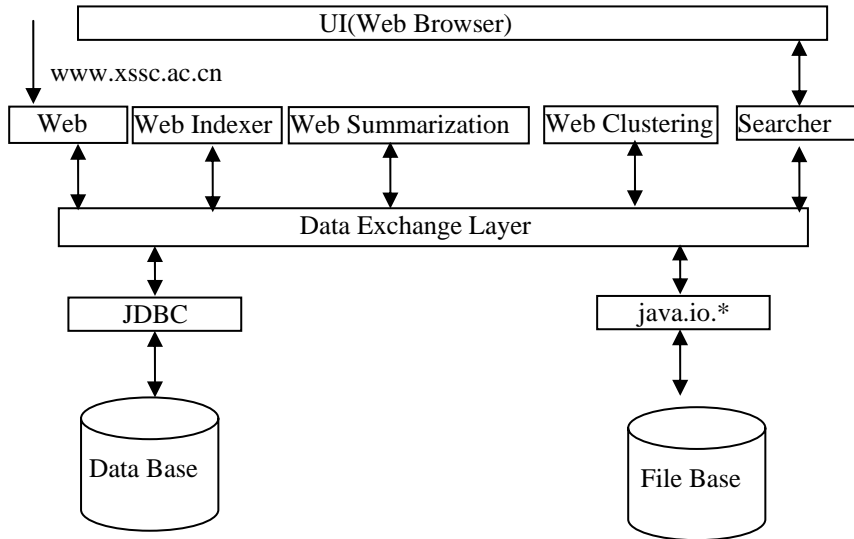
Fig. 2. System overview of web text mining

Fig. 3. The structure of XSSC web page

4.4 Web text summarization on XSSC

Here, the ICTCLAS [9] is employed to make segmentation for the extracted pure texts (also as Web texts alternately) in Section 4.3 to individual words. Only the nouns and substantive expressions were retained as the keywords candidates for each Web page. Then the 15% highest frequency keywords candidates of each pure text were selected as initial keywords collection for each Web page. Next, we combined all the initial keywords collection into an overall words collection and selected only the 5%

highest frequency words of the overall collection to construct domain words collection for all texts. Also, the domain words should be examined by experts of XSSC with heuristic method. Then we use each initial keywords collection to subtract domain words collection so as to obtain the final keywords for each Web text. The reason for this method to obtain keywords of each text is that usually there are some over-frequency words which own the high frequency in each Web text, but in fact, these words are not so significant for distinguishing these Web texts because they are ordinary words such as "science", "system", and so on.This phenomenon for texts anlalysis is well distinguished for Zipf's law [10]. Another reason is that our Chinese words-frequency distribution is not as the same as the Luhn's description.This point was discussed in detials in reference[11]. So we employed a statistical and heuristic method to obtain the keywords for each Web text. When the keywords selection is done, the classic Luhn's sentence score [12] method is employed to evaluate each sentence in pure text with each keywords collection. Finally, the predefined number of sentences at the predefined compression ratio to the whole text were extracted from each pure text and rearranged with coherence and fluency to constitute the abstract of each Web page. Here in this case, the predifined ratio is 0.25. Figure 7 is the details on processes of Web text summarization.

Figure 4：Extracted pure text
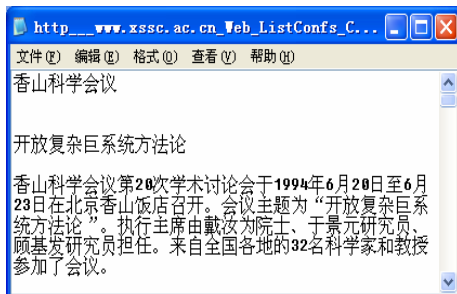

Figure 5：Extracted participants

Fig. 6. Extracted HTML tags

4.5 Web text clustering on XSSC

In Section 4.3, we selected 208 web pages as our Web content indexing and Web text sumarization samples. Here, 192 of these 208 web pages were selected for Web text clustering because there are some web pages whose content is too short and inludes too few keywords to undertake clustering. Web text cluatering is carried out as follows:

Step 1. To combine the key words of each Web text so as to generate a keywors set for all these 192 texts.

A keywords set is employed here is that we only care the occurrence of keywords instead of the frequency of each keyword in 192 Web texts. By the proccessing of this step, an overall keywords set for the 192 texts with 8392 terms is established as

$$(t_1, t_2, ..., t_{8392}).$$

Step 2. To represent the 192 texts with the combined overall keywods set using Boolean model.

By the proccessing of this step, the *i*th text of 192 texts is represented as $Doc(i)$ with a Boolean vector.

That is,

$$Doc(i) = (k_{i,1}, k_{i,2}, ..., k_{i,8392}), \ let \ k_{ij}$$

$$= \begin{cases} 1, \ if \ term \ j \ is \ existing \ in \ ith \ text \\ 0, \ if \ term \ j \ is \ not \ existing \ in \ ith \ text \end{cases}$$

Step3. To calculate the similarities using cosine included angle among 192 texts.

That is, if let s_{ij} is the similarity between $Doc(i)$ and $Doc(j)$,

then $s_{ij} = \dfrac{Doc(i) \Box Doc(j)}{\left|Doc(i)\right|\left|Doc(j)\right|}.$

Step4. To represent each text with the similarities between this text and all the 192 texts.

That is, $Doc(i) = (s_{i,1}, s_{i,2}, ..., s_{i,192}).$

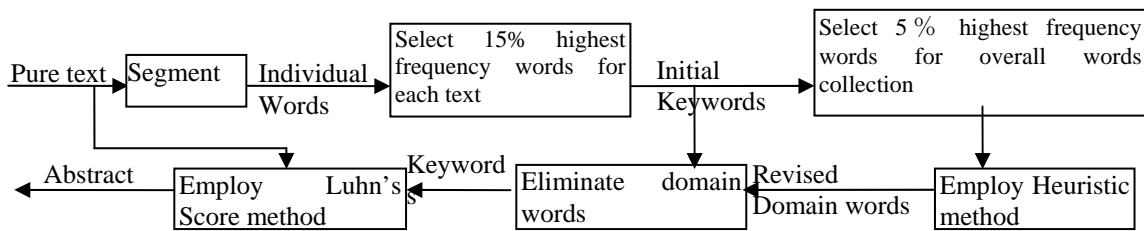*Step5. To employ the hierarchical clustering method to cluster the 192 similarity vectors.*

Fig. 7. Process of web summarization

By this method of clustering, 192 texts were clustered into 35 clusters. With the judgement of human experts, 29 of 35 clusters whose constituent texts have a common theme emerged and can be clearly labeled and 164 of 192 texts were clustered into these 29 clusters. Table 1 shows the statistics for the Web text clustering on XSSC.

Table 1. Statistics for the web text clustering on XSSC

Total number of clusters	Effective clusters	Total number of samples	Total number of samples in effective clusters	Average of precision for all clusters
35	29	192	164	0.8538

4.6 User interface to interpret the Web text mining results

Here, a user interface like search engine as Google and Yahoo is tailored to exhibit the results of Web text mining as well as for the user to retrieval Web information on XSSC. Two kinds of retrieval was provied here: one is to search for the topics with keywords and another is to search for the experts who have ever been invited for the XSSC with their names. After the user input the keywords or names in the textarea, the full text searching for thoses keywords were conducted among the downloaded and filtered local Web pages, then a feedback is returned with the hit results whose Web contents contain the words which user want to find out from the XSSC Website.

Figure 8 is an example using our information support tool to search the keyword "complexity" in XSSC, it can be seen that the information related with this keyword was retrieved elaborately from XSSC Website with 137 hit results arranged by their relevance with the input keyword. The original URL where the keyword occurred was listed with all kinds of Web text minig results for this Web page. Concretely, the mining results include following

information: the topic about the conference which is the result of information extraction conducted in Section 4.3; the highest score sentence which is derived by the Luhn score for each sentences in Section 4.4; the abstract for a Web page which is produced also by Luhn's method and a predefined compression ratio for the text to extract a number of sentences, namely Web text summarization in Section 4.4; the names of participants which is obtained by the information extraction from the Web page also in Section 4.3; the related conferences with the conference of this Web page which are generated and presented by the Web texts clustering described in Section 4.5.

As for the name searching, the user interface has provided automatic name awareness with high intelligence to retrieval the Web pages which include the query name. Figure 9 has shown the name push information based on automatic name awareness. Once a person's name is input in the textarea, the user interface can "aware" it automatically and then all the information occurred in the XSSC Website related with this person will be collected and pushed to the user timely. In brief, the pushed information includes: name, title, affiliation, conferences participated of the person and all the people who have participated same conferences simultaneously with this searched person at least twice.

With the support of user interface, information about both the specific content and the personal participants in XSSC can be obtained timely and easily.

5. Conclusion

A Chinese Web text mining process proposed in this paper includes Web crawler, Web indexer, Web summarization and Web clustering. Each module of this process involves different techniques we have discussed. The Web crawler is used to collect Web pages from Internet; the Web indexer is used to extract the features from Web pages; the Web summarization is applied to generate the abstract of Web pages automatically; the Web clustering is utilized to find the related Web pages of a Web page. After the

introduction of Chinese Web text mining process, an application example on XSSC is proposed to exhibit the power of this process. We have discussed the function of each module in the practical Web text mining by the implementation of the Web text mining on XSSC. Also, a user interface is presented to interpret the result of the Web text mining results.

Other contributions in this paper are the improvement of Luhn's key words selection method and the clustering method for the Boolean vectors.

It is should be pointed out that our Web text mining is an application part of our GAE(Group Argumentation Environment) [13]. Our research motivation is that knowledge creation would be

enhanced with the large and supervised information support. In fact, our research is accepted by XSSC administration authority [14].

However, our research is in the initial step and needs much more improvements. In the technical aspects, the need to improve spider algorithm to make it more efficiently, to improve our indexing to make it more flexible and to revise the extraction algorithm and to introduce semantic Web to avoid the deficiency of Luhn's pure statistical score method are all our further directions. In application, we only apply our Web text mining on XSSC and need to extend this application to other fields such as BBS, news web site and so on.

Fig. 8: Results by full text searching

Fig. 9: Name searching in XSSC

References

[1] N. K. Liu, W. D. Luo, M. C. Chan, "Design and Implement a Web News Retrieval System". R. Khosla, R. J. Howlett, and L. C. Jain (eds.): *Knowledge-Based Intelligent Information & Engineering Systems* (proceedings of KES'2005, Part III), LNAI 3683, Springer, 149-156, 2005.

[2] H. Chen, "High-Performance Digital Library Classification Systems: An Experiment in Medical Informatics", *Proceedings of the Third International Asian Digital Library Conference*, Seoul, Korea, 188-195, 2000..

[3] T. Nakada, S. Kunifuji. "Subgroup Discovery among Personal Homepages". G. Grieser, Y. Tanaka, & A. Yamamoto (Eds.): *Discovery Science (proceedings of the 6th International Conference, DS 2003)*, LNCS 2843, Springer, 385-392, 2003.

[4] L. Li, et al. EGA: "An Algorithm for Automatic Semi-structured Web Documents Extraction", Y. Lee et al (Eds): *Database Systems for Advanced Applications (proceedings of the 9th International Conference, DASFAA 2004)*, LNCS 2973, Springer, 787-798, 2004.

[5] R. Fieldman, I. Dagan. "Mining Text Using Keyword Distribution". *Journal of Intelligent Information Systems.* 10(3): 281–300, 1998.

[6] H.C. Yang, C. H. Lee. "A text mining approach for automatic construction of hypertexts". *Expert System with Applications.* 29(4): 723-734, 2005.

[7] F. J. Ren. "Automatic abstraction important sentences". *International Journal of Information Technology & Decision Making.* 4(1): 141-152, 2005.

[8] Salton G, Wong A, Yang C S. "A Vector Space Model for Automatic Indexing". *Communication of the ACM*. 18: 613- 620, 1995.

[9] Chinese Lexical Analysis System: ICTCLAS. Online: http://nlp.org.cn/~zhp/ICTCLAS/codes.html

[10] R. F. Cancho, R. V. Sole. "Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited". Online: http://edu.qiji.cn/eprint/abs/1289.html

[11] W. Zhang. *Information support tool based on Web Text mining and its application.* Master thesis, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, 2006.(in Chinese)

[12] H. P. Luhn. "The Automatic Creation of Literature Abstracts". *IBM journal of research and development*. 2(2): 159-165, 1958.

[13] X. J. Tang, Y.J. Liu, W. Zhang. "Computerized Support for Idea Generation during Knowledge Creating Process". R. Khosla, R. J. Howlett, and L. C. Jain (eds.): *Knowledge-Based Intelligent Information & Engineering Systems* (proceedings of KES'2005, Part IV), Lecture Notes on Artificial Intelligence, Vol.3684, Springer-Verlag, Berlin Heidelberg, 437-443, 2005.

[14] Y. J. Liu, X. J. Tang, W. Z. H. Li. "A Preliminary Analysis of XSSC as Transdisciplinary Argumentation". Liu Sifeng, Zhou Dequn, Zhang Yang and Li Hongwei (eds.): *New Development of Management Science and Systematic Science* (Proceedings of The 8th Youth Conference on Management Science and System Science), Press of Hehai University. Nanjing China, (5),35-40, 2005.(in Chinese)