Gu J F, Nakamori Y, Wang Z T and Tang X J (eds.)
Knowledge and Systems Sciences: toward Knowledge Synthesis and Creation
Proceedings of KSS2006, LNDS 8, Global-Link, September 22-25, 2006, Beijing, China

WEB TEXT MINING ON XSSC

Wen ZHANG Xijin TANG

Institute of Systems Science, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100080 P.R.China
email: {zhangwen, xjtang}@amss.ac.cn

Abstract

XSSC (Xiangshan Science Conference) is famous for its directive role of Chinese basic scientific research by inviting senior experts in research fields to express their opinions concerning their own researches. In this paper, we proposed a Chinese Web text mining process based on traditional information retrieval and extraction as well as Internet techniques. Then we conducted this Web text mining process on XSSC and the results of this process are presented. 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. Web mining is employed to improve our access ability by providing its users with useful, timely and comprehensive information support from Web. 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 final 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 and the other is focused on the algorithms or the improvements of some Web text mining methods. 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 may not be applied in other fields because of lack of feasibility. The research of Web text mining technique have not synthesized the Web features of texts and the text mining techniques effectively resulting it being separated into Web representation and text mining instead of the research on the combination of these two sides.

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. So there would be more preprocess on Web texts than pure texts. Figure 1 is a basic Chinese Web text mining process. The function of each module is as follows.

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. So there would be more preprocess 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 download Web pages from Internet. Here, the traditional spider detecting algorithm showed in the following part with two parameters as the seed URL and the depth for crawling is employed.

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 previously downloaded Web page and saved into local database.

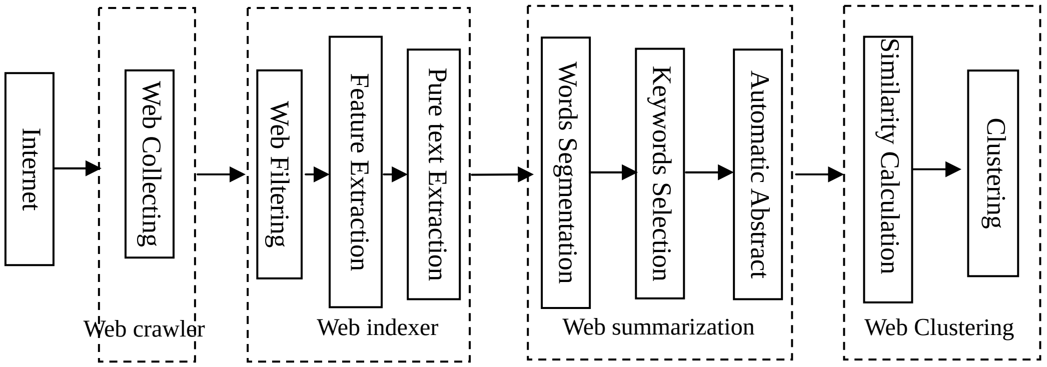


Figure 1. A basic Chinese Web Text mining process

Secondly, the Web pages at the newly parsed URLs are 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 some cases, restrictions are set for the crawler's downloading by evaluate the URLs parsed 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 checking the parsed URLs continuously.

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 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. 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 test of Web page and rearranges the extracted sentences 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 which have the most resolving power for distinguishing one 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 by 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 here to cluster the Web pages without supervision according to the similarities between them. Firstly, it can bring its user a whole grasp of the Web information by clustering the Web pages into some clusters. Secondly, it can provide user with the rest related Web pages of some one 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. Generally, 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 care the occurrence of a term in a text while ignoring the frequency. 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 in section 3, a practical Web text mining was conducted and implemented on XSSC Website (www.xssc.ac.cn). Figure 2 is the system over view of the Web text mining process 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 work of Web crawler, we obtained 646 Web pages from the XSSC Website.

4.3 Web Web indexing on XSSC

In this step, those Web pages whose contents are concerned general description of XSSC were eliminated and we selected only the conference details Web pages because they contained the overall information about a certain conference such as conference topic, the utterance records of experts and the conferences participants. After the Web filtering, 208 Web pages whose contents were within our mining interest were retained for latter Web text mining processing.

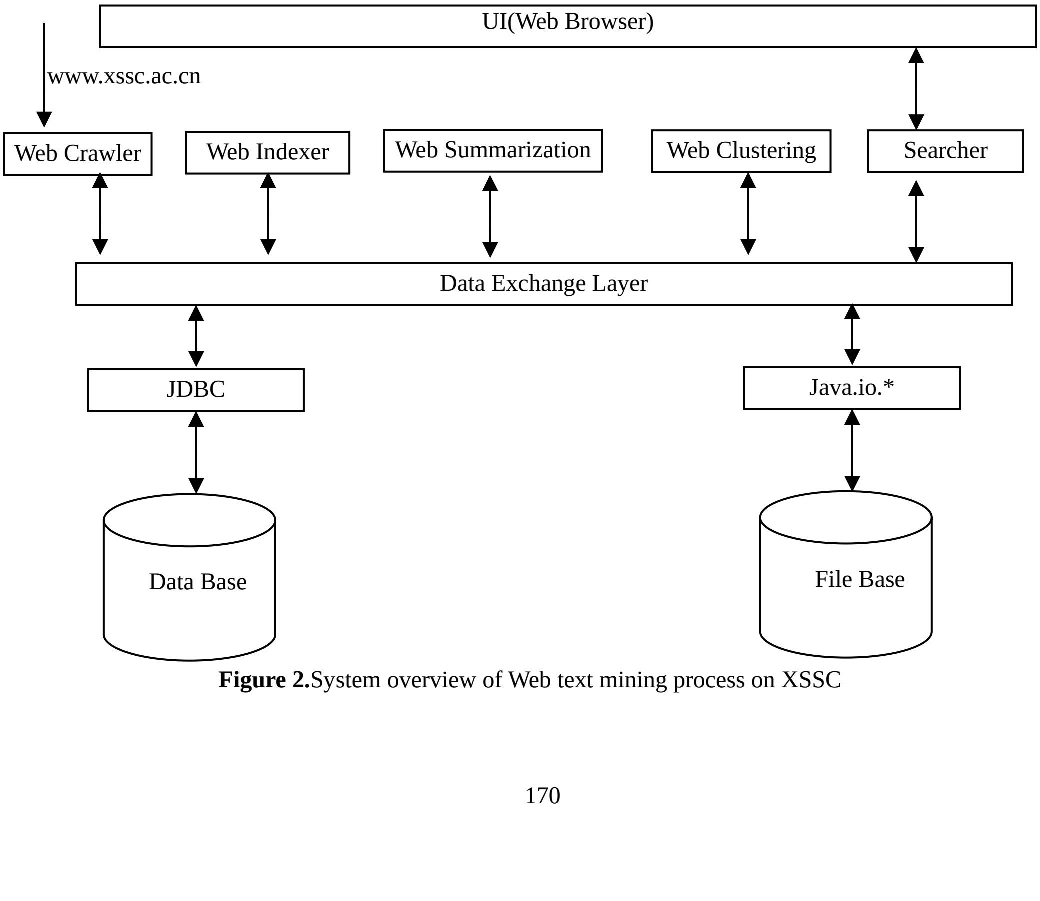
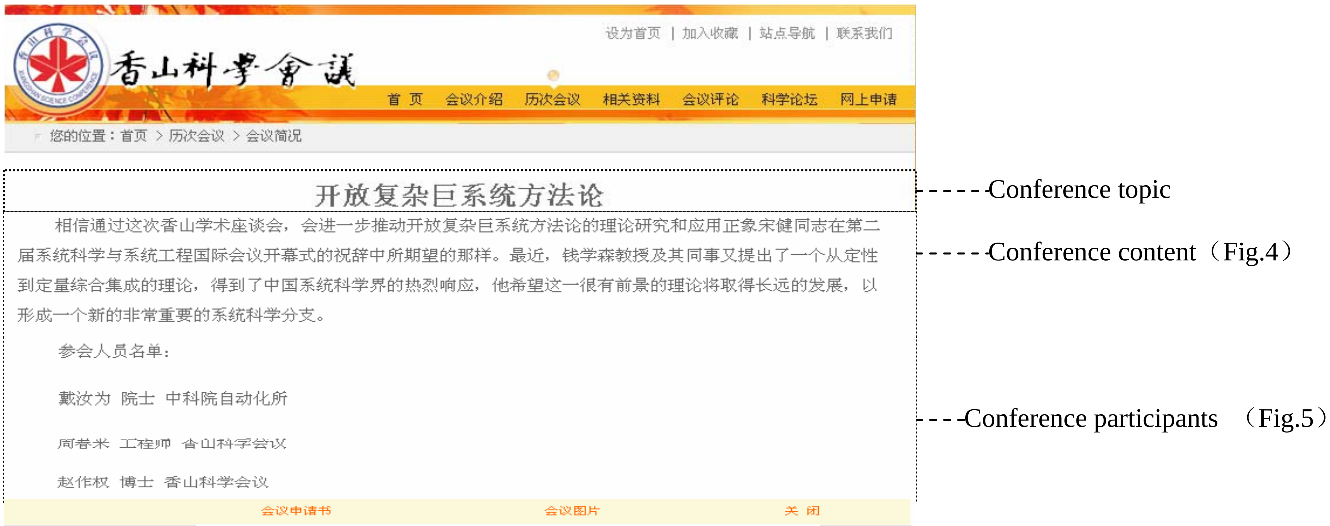


Figure 2. System overview of Web text mining process on XSSC

Figure [3] is a typical Web page from XSSC Website about conference details.It can be seen that there is certain kind of structure existing in the Web page. By identifying this kind of structure, we can extract the information about conference topic, conference content as well as the information about the conference participants. Figure [4-6] are the extraction results from the XSSC Web page.


设为首页 | 加入收藏 | 站点导航 | 联系我们
香山科學會議
首 页
会议介绍
历次会议
相关资料
会议评论
科学论坛
网上申请
您的位置：首页 > 历次会议 > 会议简况
Conference topic
Conference content（Fig.4）
Conference participants （Fig.5）
会议申请书
会议图片
关 闭

开放复杂巨系统方法论

相信通过这次香山学术座谈会，会进一步推动开放复杂巨系统方法论的理论研究和应用正象宋健同志在第二届系统科学与系统工程国际会议开幕式的祝辞中所期望的那样。最近，钱学森教授及其同事又提出了一个从定性到定量综合集成的理论，得到了中国系统科学界的热烈响应，他希望这一很有前景的理论将取得长远的发展，以形成一个新的非常重要的系统科学分支。

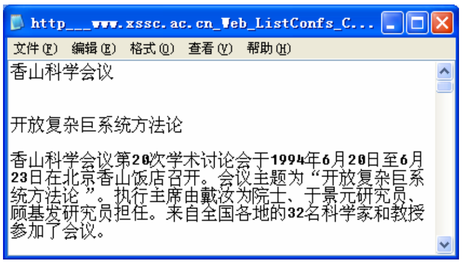
参会人员名单：

戴汝为 院士 中科院自动化所

周春米 工程师 香山科学会议

赵作权 博士 香山科学会议

Figure 3: The structure of XSSC Web page

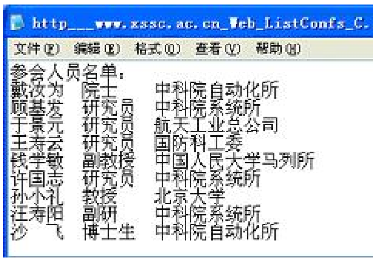

http___www.xssc.ac.cn_Web_ListConfs_C...
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

香山科学会议

开放复杂巨系统方法论

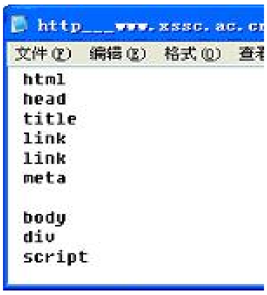
香山科学会议第20次学术讨论会于1994年6月20日至6月23日在北京香山饭店召开。会议主题为“开放复杂巨系统方法论”。执行主席由戴汝为院士、于景元研究员、顾基发研究员担任。来自全国各地的32名科学家和教授参加了会议。

Figure 4: Extracted pure text


http___www.xssc.ac.cn_Web_ListConfs_C.
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
参会人员名单：
戴汝为   院士     中科院自动化所
顾基发   研究员   中科院系统所
于景元   研究员   航天工业总公司
王寿云   研究员   国防科工委
钱学敏   副教授   中国人民大学马列所
许国志   研究员   中科院系统所
孙小礼   教授     北京大学
汪寿阳   副研     中科院系统所
沙  飞   博士生   中科院自动化所
```

Figure 5: Extracted participants



```
html
head
title
link
link
meta

body
div
script
```

Figure 6: Extracted HTML Tags

4.4 Web text sumarization on XSSC

Here, the ICTCLAS [9] is employed to segment the extracted pure texts (also as Web texts) 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 of 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. Another reason is that our Chinese words-frequency distribution is not as the same as the Luhn's description [10]. 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 [10] method is employed to evaluate each sentence in pure text with each keywords collection. Finally, the predefined number of sentences were extracted from each pure text and rearranged by coherence and fluency to constitute the abstract of each Web page. Figure 7 is the detailed process of Web text summarization.

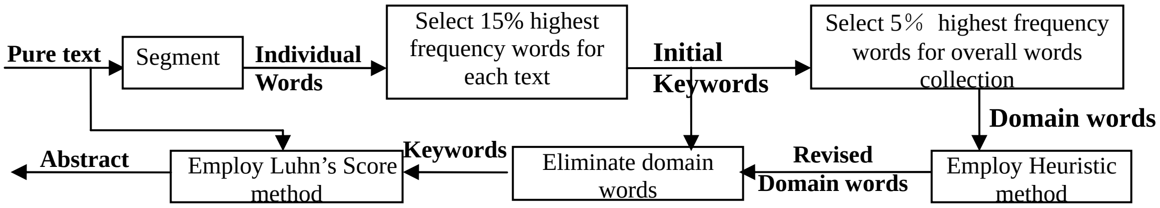


Figure 7：Process of Web Summarization

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

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

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

That is,

$$\mathrm{Doc}(i) = (k_{i,1}, k_{i,2}, \ldots, k_{i,8392}),\ \text{let } k_{ij} = \begin{cases} 1, \text{ if term j is existing in } i\text{th text} \\ 0, \text{ if term j is not existing in } i\text{th text} \end{cases}$$

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

$$s_{ij} = \frac{Doc(i) \bullet Doc(j)}{|Doc(i)| \times |Doc(j)|}.$$

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

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

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

By this method of clustering, 192 texts were clustered into 35 clusters. From the view of author, 29 of 35 clusters whose constituent texts have a common significant theme are clear to understand and 164 of 192 texts were clustered into these 29 clusters. Table 1 shows the statistics for the Web text clustering on XSSC.

4.6 User interface to interpret the Web text mining results

Here, a user interface like search engine is tailored to interpret the results of Web text mining as well as for the user to retrieval Web information on XSSC.By the full text searching in the downloaded local Web pages, a feedback is brought forth by the user interface with the hit results whose Web contents contain the words which user want to find out from the XSSC Website.

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

It can be seen from figure 8 that each hit result is made up of six parts: the original URL of the hit Web page, the highest Luhn's score sentence in Web text, the hyperlink of the automatic abstract generated by Web text summarization in section 4.4, the hyperlink of participant generated by extraction in section 4.3 and the related results generated by Web text clustering in section 4.5.

5. Conclusion

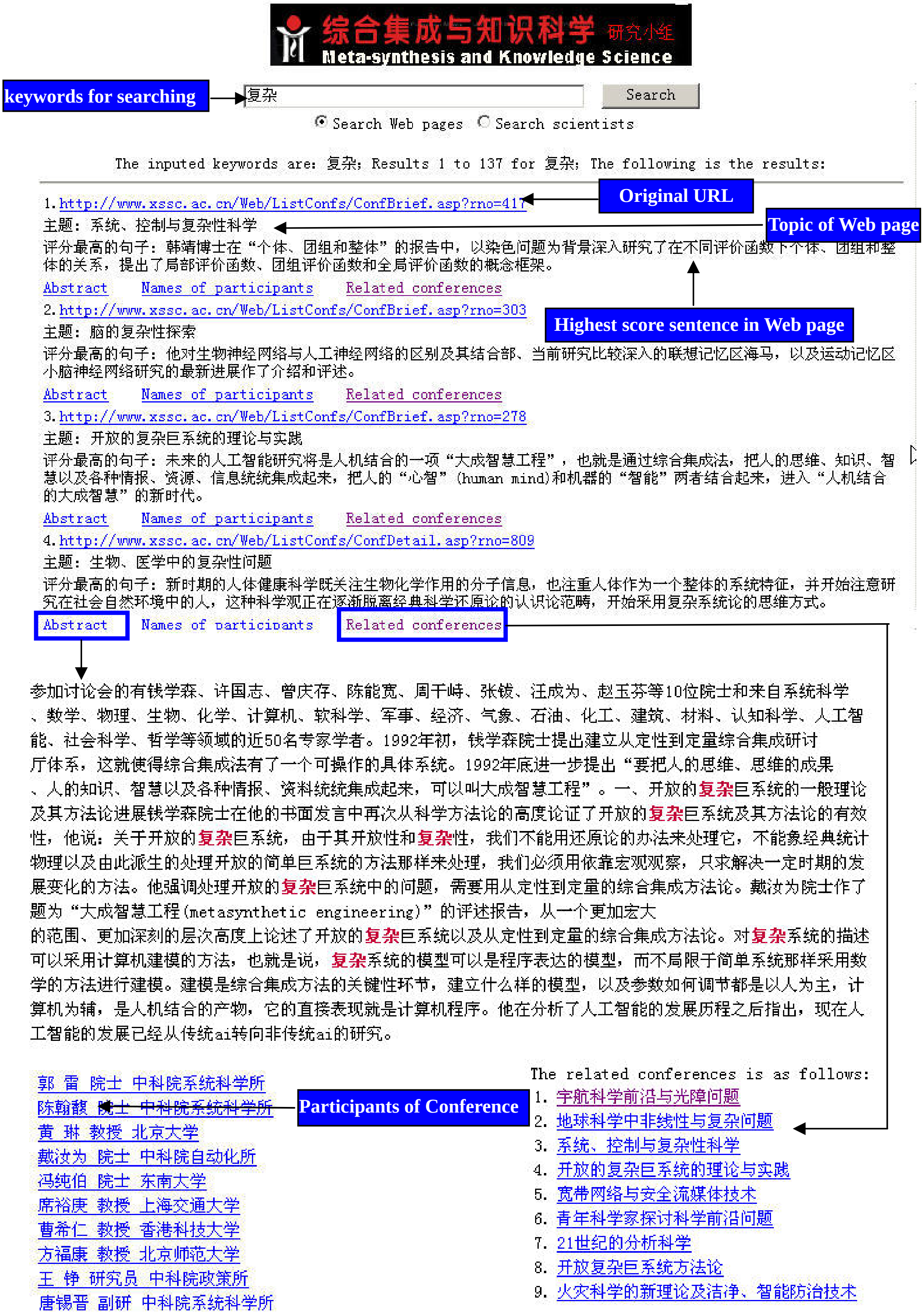
A Chinese Web text mining process proposed in this paper includes Web crawler, Web indexer, Web summarization and Web clustering. Each module of this process involves different techniques we have discussed. The Web crawler is used to collect Web pages from Internet; the Web indexer is used to extract 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)[11]. 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[12].

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 it's application to other fields such as BBS, news web site and so on.

Acknowledgements

This work is supported by national Natural Sciences Foundation of China (Grant No. 70571078 & 70221001).


综合集成与知识科学 研究小组
Meta-synthesis and Knowledge Science
keywords for searching
Search
Original URL
Topic of Web page
Highest score sentence in Web page
Participants of Conference

复杂

● Search Web pages ○ Search scientists

The inputed keywords are: 复杂; Results 1 to 137 for 复杂; The following is the results:

1. http://www.xssc.ac.cn/Web/ListConfs/ConfBrief.asp?rno=417
主题：系统、控制与复杂性科学
评分最高的句子：韩靖博士在“个体、团组和整体”的报告中，以染色问题为背景深入研究了在不同评价函数下个体、团组和整体的关系，提出了局部评价函数、团组评价函数和全局评价函数的概念框架。
Abstract　Names of participants　Related conferences

2. http://www.xssc.ac.cn/Web/ListConfs/ConfBrief.asp?rno=303
主题：脑的复杂性探索
评分最高的句子：他对生物神经网络与人工神经网络的区别及其结合部、当前研究比较深入的联想记忆区海马，以及运动记忆区小脑神经网络研究的最新进展作了介绍和评述。
Abstract　Names of participants　Related conferences

3. http://www.xssc.ac.cn/Web/ListConfs/ConfBrief.asp?rno=278
主题：开放的复杂巨系统的理论与实践
评分最高的句子：未来的人工智能研究将是人机结合的一项“大成智慧工程”，也就是通过综合集成法，把人的思维、知识、智慧以及各种情报、资源、信息统统集成起来，把人的“心智”(human mind)和机器的“智能”两者结合起来，进入“人机结合的大成智慧”的新时代。
Abstract　Names of participants　Related conferences

4. http://www.xssc.ac.cn/Web/ListConfs/ConfDetail.asp?rno=809
主题：生物、医学中的复杂性问题
评分最高的句子：新时期的人体健康科学既关注生物化学作用的分子信息，也注重人体作为一个整体的系统特征，并开始注意研究在社会自然环境中的人，这种科学观正在逐渐脱离经典科学还原论的认识论范畴，开始采用复杂系统论的思维方式。
Abstract　Names of participants　Related conferences

参加讨论会的有钱学森、许国志、曾庆存、陈能宽、周干峙、张钹、汪成为、赵玉芬等10位院士和来自系统科学、数学、物理、生物、化学、计算机、软科学、军事、经济、气象、石油、化工、建筑、材料、认知科学、人工智能、社会科学、哲学等领域的近50名专家学者。1992年初，钱学森院士提出建立从定性到定量综合集成研讨厅体系，这就使得综合集成法有了一个可操作的具体系统。1992年底进一步提出“要把人的思维、思维的成果、人的知识、智慧以及各种情报、资料统统集成起来，可以叫大成智慧工程”。一、开放的复杂巨系统的一般理论及其方法论进展钱学森院士在他的书面发言中再次从科学方法论的高度论证了开放的复杂巨系统及其方法论的有效性，他说：关于开放的复杂巨系统，由于其开放性和复杂性，我们不能用还原论的办法来处理它，不能象经典统计物理以及由此派生的处理开放的简单巨系统的方法那样来处理，我们必须用依靠宏观观察，只求解决一定时期的发展变化的方法。他强调处理开放的复杂巨系统中的问题，需要用从定性到定量的综合集成方法论。戴汝为院士作了题为“大成智慧工程(metasynthetic engineering)”的评述报告，从一个更加宏大的范围、更加深刻的层次高度上论述了开放的复杂巨系统以及从定性到定量的综合集成方法论。对复杂系统的描述可以采用计算机建模的方法，也就是说，复杂系统的模型可以是程序表达的模型，而不局限于简单系统那样采用数学的方法进行建模。建模是综合集成方法的关键性环节，建立什么样的模型，以及参数如何调节都是以人为主，计算机为辅，是人机结合的产物，它的直接表现就是计算机程序。他在分析了人工智能的发展历程之后指出，现在人工智能的发展已经从传统ai转向非传统ai的研究。

郭 雷 院士 中科院系统科学所
陈翰馥 院士 中科院系统科学所
黄 琳 教授 北京大学
戴汝为 院士 中科院自动化所
冯纯伯 院士 东南大学
席裕庚 教授 上海交通大学
曹希仁 教授 香港科技大学
方福康 教授 北京师范大学
王 铮 研究员 中科院政策所
唐锡晋 副研 中科院系统科学所

The related conferences is as follows:
1. 宇航科学前沿与光障问题
2. 地球科学中非线性与复杂问题
3. 系统、控制与复杂性科学
4. 开放的复杂巨系统的理论与实践
5. 宽带网络与安全流媒体技术
6. 青年科学家探讨科学前沿问题
7. 21世纪的分析科学
8. 开放复杂巨系统方法论
9. 火灾科学的新理论及洁净、智能防治技术

Figure 8: Results by full text searching

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, 2005, 149-156.

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, 2000, 188-195.

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, 2003, 385-392.

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, 2004, 787-798.

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

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

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

8 Salton G, Wong A, Yang C S. A Vector Space Model for Automatic Indexing. Communication of the ACM. 1995.18: 613- 620.

9 http://nlp.org.cn/~zhp/ICTCLAS/codes.html

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

11 Y.J. Liu, X. J. Tang, 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, 2005, 437-443.

12 Y.J. Liu, X. J. Tang, Z. H. Li. A Preliminary Analysis of XSSC as Transdisciplinary Argumentation. S. F. Liu et al, (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, 2005(5), 35-40. in Chinese.