Research of the Chinese Meta-Search Engine Model Based on Intelligent Agent

WANG Hao-ming FENG Bo-qin
School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, 710049, P.R.China
wanghm@mail.xjtu.edu.cn bqfeng@mail.xjtu.edu.cn

Abstract

This paper discussed the disadvantages of the current search engines and built a meta-search engine. The system constructed the vector set, which composed of the vectors of each domain, by using the materials hading been classified manually. When the user visited the Internet, the historical pages were recorded and downloaded. The system extracted the information from the pages and constructed the vector of every page. The distance between the page vector and the domain vector was calculated. The more the distance the more the relativity. The personalized model of the given user was built by the relativity. Due to the limit of the dictionary, the method of the word segmentation called "a domain independent dictionary free lexical acquisition model for Chinese document" was used.

In this paper, some key technologies related to the system were discussed, such as extracting the class keywords, classifying methods, and the algorithm of the page value calculating. In the last paragraph, the future works, such as the storing of semi-structure information, and the calculating of the web page's value were mentioned.

Keywords: Intelligent agent, Personalization, Personalized Information Retrieval, Class keywords, Meta-search engine

1. Introduction

Due to the rapid development of the World Wide Web, some features of the Internet, such as mass, semi-structure, have become drawbacks in the using of the information in Internet widely. They are [1]:

- Multi-type information coexist and locate in many sites of the WWW, and
- The structure of the information in every site is comparatively stable, whereas the validity and the reliability of the information in each of the site are changeable from time to time, and
- Due to the availability and the dynamic of the information, it maybe fuzzy or wrong.

As the information hasn't been sorted in a special order, it will be difficult to find the information the users need. That means the user cannot utilize the Internet resource effectively.

There are two expectations in information retrieval, the precision and the recall. The precision shows the degree of the feedback results case to the user's needs. The recall shows the percentage of the feedback results to the total records, which is related to the user's needs. A search engine must decide what kinds of sources should to be queried, how to modify the query expression the user submitted in order to utilize the underlying search engines more, and how to order the feedback results. Some search engines allow user to influence one of these decisions, but not all three. The user needs an efficient and intelligent system for information retrieval [2][15].

In order to implement the aims, the artificial intelligence (AI) has been suggested. Many types of IR models have been put forward. They can be divided into two items, the one is based on the large scales machine learning, and the other is based on the intelligent personalization. Intelligent agent technology is an efficient way to implement those systems. This paper described the architecture of Chinese information retrieval model by the intelligent agent, and introduced the workflow of the model. The agents gathered the historical pages the user had visited, and built the personalization model with content of those pages.

1.1 The agent paradigm

The intelligent agent (IA) is a set of the programs, which can gather the needs of the user, hand them to the current search engines, and sort out the feedback

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05)
1550-445X/05 $20.00 © 2005 IEEE

results. IA has its unique advantages, such as flexibility, activity and collaboration. In this research area, there are some relevant subjects or systems, such as Watson, built by Northwestern University, WebWatcher, built by Carnegie Mellon University, and JITIRs, built by MIT, etc.

This kind of system is composed of two modules, the agent module and the automatic-learning module. The agent module is used for the interaction between the user and the Internet, and it stored the historical pages, which the user had visited on the Internet. The automatic-learning module visits the pages that had been downloaded before, and summarizes the habits of the user in order to refine the interaction. The agent waits for the requisition of the Internet explorer in a looping way. When the requisition arrives, the agent selects the records from the local database or submits the query to the search engines. Each record is marked by relativity. And the record with the highest relativity is arranged on the top of the queue.

There are two key technologies in this kind of system at least. They are:

- **Keywords extracting**

The keywords can be extracted by the way based on the keywords appearance frequency in the documents or based on the sentence meaning analyses. The first method is used widely.

- **Results clustering and analyzing**

The results, which the current search engines returned, include many repeat records. The same content displays in many web sites. It causes the returning results low value though there are a large amount of records. The clustering and analyzing technologies are used to select the most value record from the feedback results for the user.

Actually, depend on the quality of the initial query, it will be possible that many documents return but few of them is relevant. The initial query expression may be a Boolean expression, which consisted of conjunction or disjunction keywords, or a direct question.

1.2 The disadvantage of current IR systems

Current IR systems have some disadvantages, they are:

- They cannot follow the change of the user's interest in non-personalization mode. They provide the same mode for all kinds of users no matter what domain the question belongs to. The results may make the user fuzzy. One of the important reasons is that the current systems cannot provide a convenient way for user to express his/her needs.
- They do not combine the advantages of the personalization mode and the centralized mode's. Current IR systems either focus on the large scale of information or the specific field knowledge. They haven't combined the advantages of both of them.
- The way of the interaction between the user and the IR systems is poor. It is difficult for user to express his/her needs exactly. Current IR systems need to interactive with the user many times in order to make the result more suitable for the needs of the user. It may be more suitable for different kinds of user with different ways.
- Current IR systems cannot follow the changes of the information source. The IR systems guide the user to reach the information destination he/she wanted online. It may guide the user to the websites he/she visited before. And the information may be no use for the user now. The IR systems cannot notify the user when the information has been updated.

It should be pointed that most of the current IR systems are built for searching the English information. Due to the difference between the English and the Chinese language, a Chinese user cannot locate the information easily on Internet by using current IR systems. How to improve the methods of Chinese IR has become an important factor to exert the preponderance of the Chinese information.

This paper discussed the Chinese meta-search engine based on the intelligent agent. First, it analyzed the web pages the user had visited and concluded the habits of the user and, second, it summarized the information of the user and stored it to the local server and, third, it classified the user and the information by comparability. The goal was to provide the user with the personalization service.

The paper was organized as follows: Section 2 described the architecture of the new system. Section 3 discussed the key technologies of the system involved. Section 4 and Section 5 concluded the paper and outlined the future research directions.

2. Architecture of the system

The goal of this system is to help the users to find the information on Internet easily and quickly. It requests the system should process the information resource independently, gather the information the users are interested in, filter the repeating information, wipe off the useless information, and store the

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05)
1550-445X/05 $20.00 © 2005 IEEE

information in the local database. This system is feasible, friendly, adaptable, and transplantable.

This system would not take the place of Yahoo or Google, it would be an entrance of personalized search engine. With the interaction between the user and the system, it will gather the personalized information of the user. The system includes three parts: gathering part, classifying part and issuing part. Figure.1 shows the architecture of the system.

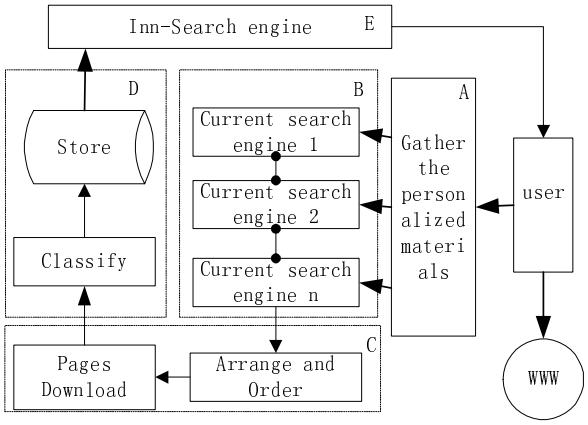

Figure 1. **The architecture of the system**

The user can either browse the Internet directly without using the personalized system or he/she can visit the system firstly in order to get the personalized service. The architecture of the system is described in details as follows:

Module A is used to analyze the user's habit automatically. At the beginning, the system needs analyze the web pages the user visited, build the personalized model, and write the information to the database in the local server. When the user logon next time, the system will provide the personalized service in order to guide the user to destination quickly.

Module B is a set of the current search engines. Module A accepts the inputs of the user and interrupts them, and then sends the information to module B. The search engines will retrieval the relevant web pages on the Internet and feedback them to module C.

Module C gets the results from module B, and analyzes them. If the same pages repeated in many sites, one of them would be hold and the others would be ignored. The page-fetching part of this module is most important. It should calculate the page's value and decide whether to download the page or not. It is easy obviously that all of the pages are downloaded and stored in the local server before the pages' value are calculated. This method will increase the network traffic dramatically. It may be fine if the page-fetching part can calculate the pages' value in the remote server, but it will impropriate the remote server's resource. Not all of the web sites allow agent to use their resource without authorization. There is not exclusive answer about which one is better. Based on the experiment experience, it may not be a good way to download all pages and then calculate the value because of the large number of the repeating web pages.

Module D is used to classify and store the information, which module C has downloaded. Two parts of the information should be stored, the contents and the hyperlinks. The classification is the key part of this module. The information, which stored in the local server, should be updated according to the remote web sites. The useful information for the given user is always ranked and arranged on the top of the queue in order to be found easily. In this system, not only the structure data but also the semi-structure data should be stored. The key technology will be studied in the future.

Module E is an inn-search engine facing the users. It is different from the search engines because it will search the information in the local database only. The feedback results may be more authoritative than the current search engines'.

Above all, information retrieval includes two sides, the general and the individual. It is difficult to build a system to achieve two aims. This system focuses on plans to implement the personalized service in some domains. It will be popularized after success in experiment.

3. The key technologies

This system is composed of three parts, information collecting, classifying and storing, and issuing. The key technologies are:

- **Collecting and amalgamating the knowledge**

The first task of IR for the users is to express their needs correctly. Because of the shortage of the tools, which the search engines can provide, the users can describe their needs faintly. This method caused the amount of feedback documents was large but few of them were relevant. This system analyzes the user's habits via the auto-learning from the historical pages in order to understand the users' needs correctly and clearly.

- **Changing and issuing the knowledge**

Most of the knowledge on Internet is non-structure or semi-structure. It is different from the data stored in the local relational database. Most of the non-structure and semi-structure data are organized as the natural language model. Those data should be converted before they are stored in the relational database, after that, they could be shared and utilized effectively.

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05)
1550-445X/05 $20.00 © 2005 **IEEE**

- **Demonstrating and applying the experiment platform**

The experiment platform is the base of the knowledge integrating with the processing technologies of all kinds of the non-normalization information. The user can index, download, collect, classify, convert and issue the non-normalization and normalization data over the experiment platform.

3.0 The SVM method

There are three methods to classify the web pages or the documents. They are Support Vector Machine (SVM), method based on the sequence, and method based on the N-Gram. The SVM is used widely.

SVM is a learning machine, which is based on the statistical learning theory. It performs binary classification and regression estimation tasks. SVM non-linearly map its n-dimensional input space into a higher-dimensional feature space. In this high-dimensional feature space a linear classifier is then constructed using quadratic programming. This latter step could potentially be very costly. SVM makes use of various kernel methods to optimize the calculation of inner numerical products. The implementation of the system performs optimizations to reduce-dimensionality of the space in order to make SVM feasible in large-dimension domains [15].

In order to determine which the category the pages or the documents are belong to, they should be word segmented. The words appeared in the documents composed a higher-dimensional feature space. The difference of the feature space between the given document and the stand document can be calculated.

The class C, which was the element of the stand classes set, could be expressed as: C=(t_1 , t_2 … , t_k), t_i(i=1,2…k) was the feature words of the class C. The $\omega_{ci} \in (0,1)$(i=1,2…k) represented the weight of the feature words ci. The class C could be expressed as:

C=(t_1 , ω_{c1} ; t_2 , ω_{c2} ; … ; t_k, ω_{ck}).

The given document D could be expressed as:

D =(t_a , ω_{da} ; t_b , ω_{db} ; … ; t_r , ω_{dr}). t_i (j=1,…,r) was the feature word of the document D, and $\omega_{dj} \in$ (0,1)(j=1,2…r) represented the weight of the feature word t_i.

The similarity of the given document C and the stand class D could be calculated as:

$$similarity(D,C) = \sum_{i=1}^{k} \omega_{di}\omega_{ci}$$

and

$$similarity(D,C) = \cos\theta = \frac{\sum_{i=1}^{k} \omega_{di}\omega_{ci}}{\sqrt{\sum_{i=1}^{k}(\omega_{di})^2 \sum_{i=1}^{k}(\omega_{ci})^2}}$$

It was obvious that the more $\cos\theta$, the more similar of the C and D.

3.1 Extracting the Class keywords of domains

Due to the limit of the number of the words can be used to represent the specifically domain. The information that decided the document belonged to the given domain of the documents should be extracted. The traditional way is segmenting the documents and calculate the frequency of the words appeared. The precondition of this way works effectively is that an accurate dictionary should be set. Any dictionary includes universal version and professional version cannot involve all of the words. In Ref. [6], there were 30 percents of 15,000 words appeared in a document had not been indexed even by a 7,000 words dictionary. The different meaning caused by different word segmenting methods is the main reason of the understanding error.

This paper used the method, called “a domain independent dictionary free lexical acquisition model for Chinese document”, to extract the feature words of the documents in order to be domain and time independent. In Ref. [11], it showed that the precision was 94 percents in high frequency words processing. It can meet the request of word segmenting in Chinese information processing. The system studied the materials independently which had been indexed on manual. The model of class keywords distributing of the given domain was built automatically.

Method 1. Assume that the document Dij (i ∈ (1,…,m) represents the classifications; j ∈ (1,…,n) represents the document which belong to the given classification) is the document which has been indexed and classified to a given classification manually. Extracting the feature words from the document, and getting the Chinese character-pairs, called $k_{ij1},\ldots,k_{ijn}$. After that, calculating the frequency of every character-pair. Using the top p(p≤n) character-pairs to build the feature vector Si of the given classification. The same way was used to build the other vectors of other classifications. In the following parts, the vector Si was represented the classification i. All of the Si consist the set S($S_1,S_2,\ldots,S_m$) of the domains.

3.2 Obtaining the personalization of the users

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA’05)
1550-445X/05 $20.00 © 2005 IEEE

The main problem of obtaining the user's personalization is the uncertainty of the user's interest. There are five technologies of dealing with the uncertainty problems. They are technologies (1) based on the Bayes network, and (2) based on the collaborative filtering, and (3) based on the DST (Dempster-Shafer theory of evidence), and (4) based on the fuzzy logical, and (5) based on the machine learning.

There are advantages and disadvantages for each of the methods. They can be used in different areas for different goals. In this paper, the system uses the technology based on the keywords picking up to collect the user's personalization [14]. It analyzes the pages, which have been downloaded according to the URLs the user had visited, picks up the class keywords, and builds the personalize model according to those class keywords. This model will be refined based on the feedback of the users.

Method 2. Assume that the document Di($i \in (1,\ldots,m)$) represents the document which the user had browsed. Extracting the feature words from every document, and getting the Chinese character-pairs, called $k_{i1},\ldots,k_{in}$. After that, calculating the frequency of every character-pair appeared. Using the top $p(p \leq n)$ character-pairs to build the personalization vector Ui of the user. The cosine distance between the vector Ui and the elements of the set S can determine the interesting of the user belong to.

The log name and the domain's name which the user interested will be written to the local database. When the user logon next time, the system requests the user input the login name. If the name was found, the system will choose the records stored in the database first, and hand the keywords, which the user submits, to the search engine. The most important is that the system hand the personalized data of the user to the inn-search engine. The feedback documents will be more relative with the user's query.

It should be pointed that the personal service is not forbidden. If the user didn't need this kind of service, he/she could visit the Internet directly.

The other work of this step was classifying the documents synchronously. The algorithm could be described as:

- Extracting the class keywords from the documents D_i, named $k_{i1},k_{i2},\ldots,k_{in}$;
- Building the feature vector V_i of the document D_i with the class keywords;
- Calculating the cosine distance between the vector V_i and the elements of the set S can determine the relativity of the document D_i and the given domain. If the cosine distance has two or more close value, a threshold filter should be set up. That means the document relates to two or more domains.

3.3 Storing the semi-structure information

The web pages are different from the traditional documents. The traditional documents have clear schema, and they could be stored in relational database in a given structure. The web pages have not clear schema. They were semi-structure or non-structure. Only when the web pages are stored in the local database can they be queried expediently. They can be described by the Object Exchange Model (OEM).

This model can also be defined by means of a few primitives: it includes the atomic type, which is a base type of the meta-model, the atomic object, which corresponds to the applications of the object type to the atomic type, and the complex object, which corresponds to the applications of the object type to an unordered sequence of (atomic or complex) objects.

In the map of the OEM, the task of web query is composed of two parts: the content querying and the hyperlinks querying. According to the granularity of the information being expressed and the function of the query language, the web query language can be divided into two levels. In the first level, the minimize unit of the map is the web page, and the border is the link between the web pages. Current search engines can implement the content query, and use the database technology to implement the structure query. This method neither takes into consideration the inn-structure of the web pages nor reconstructs the query results. The representatives are WebSQL and WebLog. In the second level, the minimize unit of the map is the inn-data of the web page, and the board can be the inn-link of the web pages or the links between the pages. Some of the models provide more natural in expressing the query results. This technology will be discussed in another paper.

4. The future research work

Some of the functions of the system discussed above have been implemented. The experiment results show that the system can provide personalized service in some of the domains with satisfaction. But it cannot be called a Chinese meta-search engine yet. The following technologies should be considered in the future:

- **The algorithm of the page value calculating**

The traditional evaluating technology of the page is based on the content of the page, the words matching,

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05)
1550-445X/05 $20.00 © 2005 **IEEE**

and the frequency of words appeared on the page. In 1998, there was an algorithm to evaluate the page value based on the hyperlinks. PageRank algorithm is a case in point.

The principle of this algorithm is that if a page was cited many times, the page may be important, while a page may not be cited many times, but it was cited by an important page, it may be important as well.

In Ref. [14], the PageRank algorithm of the Google was considered, it matched the user model smoothly. In Ref. [15], two questions were put forward. One is that a page may correlative with the given subject, but may not contain the keywords of query. It made the page fail to be select by the query. The other question is that some sites may contain a lot of hyperlinks, but it does not follow that the page may be correlative with the keywords of the query but the value of the correlative is very high.

In the future, all of the problems will be studied. The goal is to build a stand of the page valued evaluation in a given domain.

- **Web pages information extracting**

This system used the SVM method to judge the classification of web pages. It was effective when the pages had one kind of schema. There are many kinds of mass, business information, and they are arranged in another way different from those pages of science and technology. For those pages, another method should be developed, which includes the information analysis, storing and issuing, too.

5. Conclusion

This paper discussed the disadvantages, which could not provide the personalized service, of the current search engines. This paper planed to add a personalized interface, which results from the historical pages the users had visited, covering the current search engines. From the user's view, it was a Chinese meta-search engine. The user could select either the meta-search engine or current search engines.

In the other sections, this paper discussed the technologies including the user's personalization building, the keywords extracting, and the information classifying. In the last paragraph, it was pointed out that some technologies, such as the storing of semi-structure information, the calculating of the web pages' value, and the web pages' information extracting, are to be studied in the future.

Acknowledgements

This work was supported by project 03JK167 of State Education Ministry of Shaanxi Province of the PRC.

References

[1] WANG Ji-Cheng, et al. State of the Art of Information Retrieval on the Web. Journal of Computer Research & Development. Vol.38，2001(2)，p187-193

[2] Claudia Raibulet, et al. Mobile agent technology for the management of distributed systems - a case study. Computer Networks Vol.34(2000) p823-830. Elsevier Science B.V.

[3] D.Gavalasa， et al. Advanced network monitoring applications based on mobile/intelligent agent technology. Computer Communications Vol. 23 (2000) p720—730. Elsevier Science B.V.

[4] Paulo marques， et al. Building Binary Software Components for Supporting Mobile-Agent Enabled Applications. Autonomous Agents and Multi-Agent Systems， Vol.5(2002)， p103-111， Kluwer Academic Publishers

[5] Paulo Mendes，et al. Session-Aware Popularity-Based Resource Allocation for Assured Differentiated Services. IEEE Communications Magazine. September 2002. p104—111

[6] Roch H. Glitho， Edgar Olougouna， Samuel Pierre. Mobile Agents and Their Use for Information Retrieval: A Brief Overview and an Elaborate Case Study. IEEE Network. January/February 2002. p34—41

[7] A.Nur Zincir，Malcolm. Object-Orientated Design of Digital Library platforms for Multiagent Environments. IEEE Transactions on Knowledge and Data Engineering， Vol.14(2002)，No.2.

[8] A.Borodin, G.O.Roberts, J.S.Rosenthal, and P.Tsaparas. Finding Authorities and Hubs From Link Structures on the World Wide Web. Proc. 10th. International World Wide Web Conference(WWW10), 2001.

[9] Harris Drucker, et al. Support vector machines: relevance feedback and information retrieval. Information Processing and Management, Elsevier Science. Vol.(38)(2002) P305—323

[10] CAI Zhi, et al. Research of Chinese Information Retrieval on Internet. Mini-Microsystems. 2003(12). P2136—2141

[11] Giacomo Cabri，et al. Agents for information retrieval: Issues of mobility and coordination. Journals of systems Architecture. Vol.46 (2000) p1419—1433， Elsevier Science B.V.

[12] Pan Chun hau, CHANG Min, WU Gang shan. Design and Implementation of a Web Page-gathering Tool. Application Research of Computers. 2002(6). P144—147

[13] DAI Liu ling, HUANG He yan, CHEN Zhao xiong. A Comparative Study on Feature Selection in Chinese Text Catagorization. Journal of Chinese Information Processing. Vol(18) 2004(1). P26--32

[14] D.Rafiei and A.Mendelzon. What is this page known for? Computing web page reputations. In 9th International World Wide Web Conference, Amsterdam, Netherlands, May 2000.

[15] R.Lempel and S.Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. In 9th International World Wide Web Conference, Amsterdam, Netherlands, May 2000.

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05)
1550-445X/05 $20.00 © 2005 **IEEE**