The Research of Web Information Retrieval based on Temporal Information

Zhe Wang [1st,]
1st. College of Information Science and Technology
Donghua University
Shanghai,China
wzhe_wz@126.com

Chenggang Xu[2nd]
[2nd]Institute of Information Technology
Henan University of Traditional Chinese Medicine
Zhengzhou, China
xuchenggang001@126.com

***Abstract*—Temporal information is an essential attribute of the web pages, such as the publish time and the content time in the web pages. However, the major search engine hasn't more view on the temporal information of web pages,and ignored the relationship between the keywords and time phrases.In this paper, we focus on the need in time phrases recognition and extracting from the web pages,and identify the closely relationship between the keywords and the temporal information. The experiment results showed that the information retrieval based on temporal information is better than the information retrieval of simple text-keyword search in the capabilities of query expression and query processing.**

Keywords-Temporal Information; Information Retrieval; Time Phrase; Time Phrases Recognition; Time Phrases Extracting

I. INTRODUCTION *(HEADING 1)*

With the rapid development of the Internet, the amount of web data increased quickly,the search engine has become an important tool of gathering information from the mass of web pages,which enhance the internet usage and adhesive of the users. Moreover, the search engine is not only the tool of information gathering,but also become the entrance application to go to the traditional web portals. According to the statistics of CCNIC28th[1],by the ending of June 2011, the users of search engine reach 386 million, the usage rate is 79.6%,and the search engine still rank the first in the internet applications.

To make the search results closer to the needs of users and satisfied with the users, the major search engines continue to improve the search function and interface display,however, the search results often return a large number of irrelevant links,and the users have to find the needed information by pages and pages,which waste the users time and energy.So how to retrieve the information efficiently become an important research topic.

Time phrase is one of the important properties of the information,many web pages contain temporal information,and many query also contain the temporal query information. The study shows that 70% of the query keywords contained time or space [2],which 7% of user queries contains implicit-time and 1.5% of the user query contains explicit-time [3,4]. Typically, a query with temporal usually contains a few keywords and one or more of the time,the keywords and temporal words have implied relationship between them to the user.However, the major search engine hasn't more consideration on the temporal information of web pages,and ignored the relationship between the keywords and time phrases,that will lost important information of web pages in the query and temporal correlation,which can't meet the user's temporal query needs ,moreover,it ignores the important role of temporal information in the web information retrieval.

Research the temporal information integrating into the information retrieval technology is a hot trend,and google also has the function of timeline search.In the researching on temporal information of web search engines, adam[5] taked into account the users not only want information with relevant but also fresh,he proposed a way to display the web pages of update frequency. Tomoyo, etc.[6] filter the keywords by analyzing the temporal relationship between the keywords,and make the search results more accurate. Andrzej[7] proposed establish the index in the information retrieval system based on temporal information,which can provide a clear and concise description for the web document.

Temporal information play the important role in the database and information system gradually,but the technology or software products in temporal information of information retrival is very rare.In this paper,we focus on the way of identificating and extracting the time phrases from the web pages,and classificating the time phrases based on the time rules,which use the different approach for identificating and extracting the different types of time phrases,finally,we select the most relevant set of the keywords and time phrases by analyzing the web content.

II. RESEARCH ON TEMPORAL INFORMATION

Studies have shown that the proportion of temporal information only after proper nouns,in the web pages,the temporal information can be explicit expressed by time words, noun time phrase, prepositional time phrases,etc. In addition, the time information can also be implicit expressed[8]. Therefore,the processing of temporal information is a very important part of natural language understanding.

The complexity of the Chinese language lead to the time expression diversity, so researchers classified the time phrases into a different perspective. According to the continuity of temporal information, time phrases can be divided into the point of time phrases and the phrase of time phrases,such as

"June 3,2012", "Yesterday", "next Friday",etc. According to the uncertainty of temporal information, time phrases can be divided into deterministic time phrases and vague time phrases,such as "the past six months", "ten days",etc. According to the relativity of temporal information, time phrases can be divided into absolute time phrases and relative time phrases,such as "July 22,2012" is the absolute time phrase,while "April 5", "yesterday afternoon" is not unique,which is the relative time phrases.Because the relative time of expression is not uniqueness, sometimes we need to transform the relative time phrases into the absolute time phrases.

For the query of time phrases,there are several ways in common query.First,query is the point of time phrase,it set the day in the time granularity, which is a query to a specific date,such as "2009-1-1" is a point time phrases.Second,query is the phrase of time phrase, it is a time range ,which may be a few days, a few months, a few years,and so on,for example,the "2011-3-1 to 2011-5-3" and "2008-2009" are all the phrase of time phrase.Third,query is the history of time phrase,which is focus on the time of historical events,such as the query is "2008 earthquake".Finally,the query is the future of time phrase,which is focus on the time of future events.

The temporal Information of web pages can be devided into two types,one is the publish time or update time from the server,the other is the content time of web pages. For the content time of web pages,it also can be devided to explicit-time and implicit-time,the explicit-time of temporal expression can be found on the calendar easily,such as "april 2011",while the implicit-time of the temporal expression can be found by contacting the context of artificial judgment on the calendar,such as "tomorrow", "two weeks later," last month",etc.

At present,the foreigner research on temporal information of information retrival is more than our state, they focus on the publish time or the paper record time in temporal information,while the state mainly focus on the spatial and temporal databases and temporal index, and little study on the temporal information,so it is a hot trend to research the temporal information on the information retrival in the state.

III. RELATED WORKS

In this paper,we did the pre-processing work on removing the irrelevant information, Chinese word segmentation,extacting the main body in web pages.Though the experiment ,we recognized the time phrases and extracted the time phrases from the web pages,and builded the most related mode of the keywords and the time phrases.

A. The Web Pages Pre-processing

The most important difference in web pages and text is that the web pages contains a lot of labels, therefore,in the web pages,first we need to remove the interference tag from web pages.The key work of this stage is removing the irrelevant information and extracting the information,and paragraph segmentation, Chinese word segmentation,etc. The output of this stage is the each paragraph, sentence, and a set of keywords and time phrases.

The main work of pre-processing in web pages is:

- Delete the interference label of web pages like the lable of <script>, ,etc,and then segmented by <p>,
, <li>,etc.Based on the divided paragraph,we divided the paragraph into sentences by using the closing punctuation of "。", ", ","！", "?",etc.In the process of pre-processing,we used the open-source tool of ICTCLAS to devided the sentence into words[9].Based on the statistics,we found the speech of the user's query mainly contains nouns, verbs, adjectives, while rarely contains adverbs, prepositions, and interjections. Therefore, in the extraction of keywords, we ignored the parts of speech as adverbs, prepositions, and interjections words, which reduced the data scale of processed.
- Based on the meta of web pages,we extracted the publish time or update time from web pages.
- We extracted the title from each web pages by the lable of <title>,and devided the title into a set of keywords,moreover, determine the time frame with the title,which is referred to be the time of the title . However,if the title does not contain the time words,we used the update time as the title time of web pages.

B. The Research Methods

In this paper,we acquired the content time of web pages by using the traditional temporal information labeled.For the extracted time phrases,we recorded the sentence of containing the content time and the position of them.

1) The Time Phrases Recognition

To find the time corresponding to the keywords,first,we need to identified the time phrase from the web pages.In this paper,we use regular expressions and chinese word segmentation to identify the time phrases.and the sentence is an integral part of the paragraph to determine the time,which is the basis work of recognize the time phrase.

The current method of phrase recognition is based on the rules and machine learning [10].Before the recognize the time phrases,we need to build the time phrases words and boundary word dictionary.Based on these rules, scanning the words of web pages.If find the time phrases,we should judge the words whether have the left boundary time or right boundary time,if have,we use the left boundary time or right boundary time to be the time's starting point or end point.If haven't,we use the time phrases as the starting point of keywords.

The mainly time phrases have several kinds,such as:

- Time phrases:such as "April", "The day before yesterday", "yesterday",etc.
- Time conjunction or restraint word: such as "until","since","as far as","in", "deadline","up to","about","around",etc.
- Time boundary words:such as "from to","since","after","before","in the meantime","prior to","when","begin","in the middle of",etc.

2) Extracting the Time Phrases from the Web Pages

Publish time or update time of web pages often as a reference time to reasoning the correct time, the publish time of web page appeared as the form of "2011-5-11 11:49",then the output result of chinese word segmentation is "2011-5-11/m 11:/nx 49/m" ,which is recognized as the numeral.To distinguish with numeral,the time phrases should be processed before word segmentation.

For the publish time of web pages,which was often appeared below the title,we can recognized the published time through the position,and its form is regular.The usually form time expression has "YYYY MM DD, YYYY-MM-DD", "YYYY.MM.DD", "YYYY/MM/DD",etc.We can identified the time phrases by the unified regular expression,and fill in the template of the "YYYY MM DD".If the publish time form is "MM DD",we should fill the year of server as the reference time.

For the content time of web pages, the main ideas is judge the web page whether contains the time word by the regular expression or word labled,and calculated the similarity between the title and keywords,select the maximum similarity of the sentence as the main time.If the web pages haven't contain the content time ,we use the published time as the main time.

The temporal information of web pages can be an explicit time expressions, such as "October 1, 2011",etc.Sometimes it may be the implicit time expressions, such as "the day after tomorrow", "two days ago,etc.For the nonstandard time,we need to reasoning the correct time corresponding on the calendar time.We build a specification rule for the time phrase of web pages,which has the words of time that can reasoning the correct time according to the different specification rules.

According to the characteristics of specification rules,it mainly has three types:

- The time phrase with displacement calculated. such as "last year" is the year reduce one year, " next month" is the month plus one month by the reference time.
- The time phrase with proprietary words.These words represent a particular time,but not fully established. The part of time phrase in the year, month, three dimensions, one or two dimensions are known,such as "Dragon Boat Festival" is "May 1",while its year also need to be calculated by the reference time.
- The time phrase with fuzzy. such as the "fall" is the month of July to September,and the "midmonth" corresponds to 10 to 19 of days.

However , the relevance of sentence and the time is not closely,but has the similarity with the adjacent sentence of no time, a few sentences together to express an event,in this case,we need to calculate the similarity with one or more sentences.In this paper,we improved the algorithm with calculate the similarity of the paragraph first,and sorting the paragraph beginning with the largest similarity,then extract the time phrases from the paragraph, calculate the similarity of a sentence context, select the time of maximum similarity sentence as the main time in web pages.

IV. CONCLUSIONS

Temporal information as an important dimension of information,it plays an important role in information retrieval.While the major information retrival did not take full advantage of temporal information and can't provide better search results and user experience.Research on the information retrieval based on temporal information can improve the search results and promote the development of web information retrieval technology.

In this paper,we used the word correction and regular expression methods to identify and extract the time phrases,and we builded the specification rules to get the correct time according to the different specification rules,and determine the relationship between the keywords and the temporal information.However,the diversity of time phrase lead to the existing rules can't cover all the time expression,in the future,we plans to expand the rule sets and analysis the the sentence with semantic level, which can improve the accuracy of temporal information.

REFERENCES

[1] http://www.cnnic.cn.

[2] M.Wechsler,The Probability Ranking Principle Revisited,Information Retrieval,2000.

[3] D.Metzler,R.Jones, F.Peng, R.Zhang,"Improving Search Relevance for Implicitly Temporal Queries,"SIGIR(2009).

[4] S.Nunes,C.Ribeiro,and G.David.,"Use of Temporal Expressions in Web Search,"In Advances in Information Retrieval,30th European Conference on IR Research,ECIR(2008)580–584.

[5] J.Adam,K.Yukiko,Katsumi Tanaka,"Temporal Ranking Of Search Engine Results,"National Institute of Information and Communications Technology,2005:43-52.

[6] K.Tomoyo,S.Kazutoshi,"A Web Search Method Based On The Temporal Relation Of Query Keywords,"Lecture Notes in Computer Science,Web Information Systems-WISE 2006:7th International Conference on Web Information Systems Engineering,Proceedings.2006,4225:4-15.

[7] B.Andrzej,"Language Model Based Temporal Information Indexing,"Lecture Notes in Business Information Processing,Business Information Systems-11th International Conference,Proceedings.2008,7:24-35.

[8] Z.Guorong,Y.Erhong,"The Recognitionon the Time Phrase of Event Class,"Chinese Information Processing Society of China, (JSCL-2005).2005:335-337,339.

[9] http://www.ictclas.org/.

[10] S.Frankr, K.Graham,P.James,"Annotating,Extracting And Reasoning About Time And Events,"Lecture Notes in Computer Science,2007,4795:1-6.