Theme Correlation algorithm based on the domain ontology in the Vertical search engine

Mingshan xie[1, a *], Yanfang Deng[2,b]

[1] College of information engineering, Haikou College of Economics, china

[2]Hainan Technology and Business College, china

[a]283921977@qq.com, [b] 271190993@qq.com

Keywords: Vertical search engine, Theme Correlation algorithm, domain ontology

Abstract. Vertical search engine research, more and more .subject-oriented vertical search has become one of the hotspots. This paper presents an ontology judgment theme relevance judgments algorithm, combined with Semantic Web technology with traditional information retrieval technology, and referenced by user's feedback, effectively improve the user‘s search precision and recall rate, verified by experiment.

Introduction

With the rapid development of the Internet, Web Information resources have grown explosively. Generic search engines cannot be good for users with specific topics. Keyword-based search engine technology also makes the lack of semantic support. Users can not describe to know what he was looking for, except to let him see things looking. All of these gave to the search engine brings new challenges. For lack of general search engines, People want to see a more targeted search engine will appear. Vertical search engine come to provide users with more effective services, professionally, meticulous way of information processing.

Vertical search engines crawls relevant fields of knowledge directly, using focused crawler. Only selectively focused crawler access to pages with the most relevant pages given topic. There are many studies focusing on the target reptiles crawl description or definition of the web pages or data analysis and filtering, and search for the URL control strategy algorithms and implementation techniques, etc. Theme focused crawler requires that the target object being searched as consistent as possible with their own search online resources in the search process. Generally speaking, there are two kinds of methods for the calculation of this consistency: One way is based on the link; another method is based on the degree of similarity or content. Similarity-based calculation is very simple and generally uses the vector space model which does not quite accurately reflect the semantic information of Web documents. Literature [5] proposed a theme based on the dynamic correlation matching algorithm, based on meta-data and web-based method for determining the semantic content determination method, drawing on the idea of Huffman coding. This algorithm is a lack of understanding of the intent of the user search, based on the input to the associated type of calculation, by analyzing the type of data to use to adapt the determination method, and has some limitations to improve accuracy. Literature [6] proposed ontology-based information retrieval model, taking into account the efficiency of knowledge representation and reasoning capabilities of description logic to build the body, the use of tableau algorithms and case containing only atomic roles equivalence relation between individual and individual set of concepts were generated sets the quotient set, resulting in a collection of semantic index entries, use of these semantic index entries to generate better reflect the semantics of documents and user information needs logical view documents and user information needs logical view; Literature [7] determine the theme of relevance, using ontology. Literature [6] and [7] are not considered user feedback, yet research emphases placed on the ontology. In fact, the user feedback to better reflect the user's search intent, close to the user's search intention is rather high accuracy. In fact, user feedback can better reflect the user's search intent. Close to the user's search intention is rather high accuracy. Literature [7] adjusts the degree of use of user feedback, through utilization of various relevant factors, to achieve the dynamic adjustment of correlation, but the lack of systematization of these factors.

Ontology, used for specific areas of the concepts and terminology to give a clear formal description, not only provides the basis for standardization for resource description, but also to provide a guarantee for the search for information more accurately. With the diversification of network information and isomerization of network database, ontology has been increasing emphasis on the computer industry. This paper proposes a new theme correlation algorithm, based on one kind of the topic ontology in the vertical search engines, used domain knowledge-based ontology advantages and combined with user‘s feedback.

Ontology-based theme correlation computation

Ontology

Ontology is a standardized description about domain concepts and relationships between concepts, which is standardized, clear, formal, shared.Ontology goal is to capture relevant knowledge in the field, provide a common understanding of the domain knowledge to identify common recognition in the field of vocabulary, and gives these words and vocabulary clear definition of the relationship between the different levels from the formal mode.Ontology based resources become relevant to the subject calculation depends, by defining a series of concepts, relationships, examples, etc. to describe the structure of the domain knowledge in this paper.

Keywords priority value set

Before calculating correlation in the subject, this article on the use of ontology weighted keyword. Keyword weights reflect the ability to keyword performance topics,in other words, the more similar keywords and topics that should be the theme for the right keyword values will be higher, and vice versa.

When users enter their own search engine search intentions, the search engines use text mining techniques to extract core words(t_0) of the theme Keywords. This paper sets ontology library that covers the core topics related search words, while setting the theme of the core words(t_0). Because Chinese has a lot of synonyms and different users enter the same meaning, but with different words, so to locate the core words（ t_0） in ontology library collection synonyms. This paper sets that Synonym Collection of the Core words（t_0）is Ct, Initial value of the core set of synonyms Ct words each word similarity weights is 1.0.

Research has shown that distance and similarity of the words and expressions are closely related. The greater the distance between two words (Dis), the lower the degree of similarity; the contrary, the distance between two words (Dis) is smaller, the greater the similarity. Here, the distance refers to the number of edges connecting the two concepts in the ontology of the shortest path.

In order to establish a correspondence between the two, satisfy the conditions:

Similarity value is 1, when the distance between two words of the value is 0. Vocabulary and themself, as well as words and their synonyms distance values are 0.

Similarity value is Infinity, when the distance between two words of the value is 1. Actually a finite set of vocabulary is limited. Actually a finite set of vocabulary is limited and the maximum distance of two terms can be calculated and expressed. Here, infinity only indicates a trend. According to the distance between words and Ct vocabulary, this paper can define the weight of non-core.

$W(\overline{Ct})$ refers to the weight of non-core vocabulary $\overline{Ct}$.

$$W(\overline{Ct})=\frac{k}{k+\text{dis}（Ct,\overline{Ct}）} \qquad （1）$$

Here, k is the adjustment factor;0<= $W(\overline{Ct})$ <= WT;WT is the threshold; In addition, because the distance of a word in the vocabulary with the Ct may not be unique, so there is any ti, ti$\in \overline{Ct}$, t0$\in$ Ct， i is a natural integer; ti is the theme in the core ontology around the word in the i-th related words. For the collection of non-core vocabulary words ti, its weight is Wti.

$$Wti=\frac{k}{k+\text{dis}（t0,\ ti）} \qquad （2）$$

Set of the correlation weights of Theme core words: WCt={1，Wt1，Wt2，……Wti……}.

Correlation calculation using user feedback

Ontology design is good, you can use it to calculate Web page theme Correlation. Correlation calculation refers to calculate whether the content and themes related to the content and define the Correlation is much, coming from a Web crawl above data after pretreatment, the transfer to the vertical search engine Correlation calculation module. Studies have shown that the performance of Web content at the time the role of subject is different, such as the title of the content is usually shows the main content of the current page, the page in bold italics indicate the author may like to emphasize that the content. To highlight this feature, the other not to put too much weight of dispersion, the page content is divided into two parts - the title and the body, when the processing time, we give a higher weight to the title of the content. Meanwhile, the page calculated by search engine is not necessarily needed by the user and there are some inaccuracies. Vertical search engine will list the page title and a brief description related to the topic in the index database. Under normal circumstances the user when the primaries will look at the title page, and then quickly scan the next brief explanatory text, decide whether he wants to enter the page through the URL. If search engine lists the title and a brief description of the current search results determine the item does not fit his search purposes ,the user is not going through the url into the page, so the user clicks on the page (ie page feedback factor) also reflect the relevance of pages and topics.

Correlation between pages and themes in a vertical search engine can be defined as

$$sim(D, p) = (l_T sim(D,T) + l_B sim(D,B))(1+r) \qquad (3)$$

Here, D defines the subject matter; p represents a page;B, T, W denote the body of the page, the page title and weights; $l_B + l_T = 1.0$; $0 < l_T, l_B < 1.0$; r is the page feedback factor, and proportional to the time the user enters the page. Users stay on the page longer, generally described, the user's interest is relatively large for this page. A page for a long time has not been a user clicks, indicates that the page is not associated with the topic. This paper defined the acceptance of the user to page is T. the value of the user's acceptance of the first page which entered into the index library is set to 0. into the index library, a long time no one click, T will decrease, the more users click in people, T will increase more. After entering the index database, if the period of time not to click on it, the value of the T of the page is reduced, the number of times a user clicks into it more, T is increased the more.

T=0. When the page was included in the index database;

T=0-t/n. Here, t is the number of days into the index database; n is the adjustment factor determined by the administrator, according to Frequent rate that customers use search engines. If the search engine has a high frequency of use, you can adjust small point n, if the search engine has a low frequency of use, you can transfer large point n.

T=0+a/m. ‘a’ is the number of page requests, in the development of vertical search engines (based on c # development), they can use application object to record number of page requests. m is an adjustment factor. If the search engine uses high frequency, they can transfer a small dot m.If the search engines use low frequency, m can be adjusted big points.

$$r = \tan^{-1} T \qquad (4)$$

Because the feedback factor r cannot be too great an impact on the Correlation of the pages and topics, this r should be limited to a certain range . the r limit (-1,1) is more appropriate.

The formula (3) further decomposition:

$$sim(D, p) = \left(l_T \sum_{t \in T} sim(D, t) + l_B \sum_{b \in B} sim(D, b) \right)(1+r)$$

$$=（l_{\mathrm{T}}\sum_{\mathrm{t}\in\mathrm{T}}\mathrm{sim}（D,\ \mathrm{t}）+l_{\mathrm{B}}\sum_{\mathrm{b}\in B}\mathrm{sim}（D,\ \mathrm{b}））（1+\mathrm{r}）$$

$$=（l_{\mathrm{T}}\sum_{\mathrm{t}\in\mathrm{T}}\mathrm{W\,t}+l_{\mathrm{B}}\sum_{\mathrm{b}\in B}\mathrm{W\,b}）（1+\mathrm{r}） \qquad (5)$$

The correlation of the title vocabulary t and themes is directly represented by the weights of t in the ontology. Likewise, the correlation of the text vocabulary and themes is directly represented by the weights of b in the ontology. The Correlation is normalized as follows:

$$\mathrm{sim}(\mathrm{D}\ ,\ \mathrm{p}\ )=\frac{l_{\mathrm{T}}\sum_{\mathrm{t}\in T}\mathrm{W_t}+l_B\sum_{\mathrm{b}\in B}\mathrm{W_b}}{lN_T+lN_B}(1+\mathrm{r}) \qquad (6)$$

Here, N_T and N_B denote the title of the page spited and the number of the text's keywords. When using the formula (6) to calculate, the ideal situation is that the page the weight of each keyword values are 1.0, then sim (D, p) = 1 + r> = 1.0. The best ideal situation is that sim(D , p) tends to 2.0.

Theme relevance judgments algorithm flowchart

The process of the topic relevance algorithm judgment in this paper has been showed in Figure 1. In the figure, "Input" is the input domain ontology, a Web page after the pretreatment, adjustment parameter k (based on experience and generally 0.7 ~ 0.9). After inputting, initialize the page weight w、title weight Wt and body weight Wb ; In the picture, Wy is is the weight of the title words y, Wz is is the weight of the body vocabulary z, W is the theme of the page relevance, Wtd shows the page related degree threshold. Calculate the weight average of N pages:

$\mathrm{W}=\sum W_{\mathrm{x}}$ / N, N represents the number of pages that have already been processed. Adjust the value of Wtd: Wtd= W, The threshold value of the feedback is a non-human intervention threshold automatic adjustment.

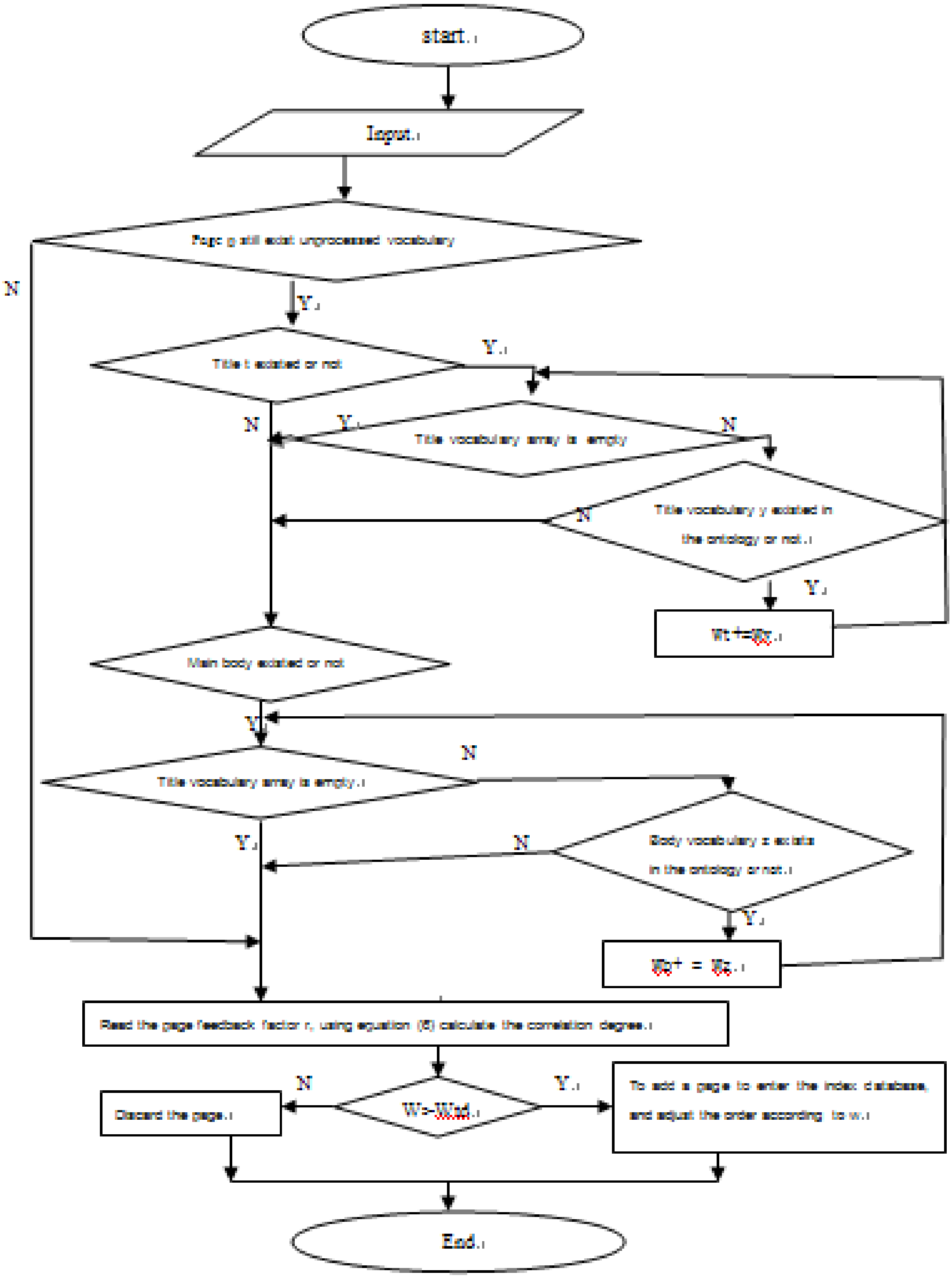
start.
Input.
Page p still exist unprocessed vocabulary
N
Y.
Title t existed or not
Y.
N
Y.
Title vocabulary array is empty
N
N
Title vocabulary y existed in
the ontology or not.
Y.
Wt+=Wy.
Main body existed or not
Y
N
Title vocabulary array is empty.
Y.
N
Body vocabulary z exists
in the ontology or not.
Y.
Wb+ = Wz.
Read the page feedback factor r, using equation (6) calculate the correlation degree.
N
Y.
Discard the page.
W>=Wnd.
To add a page to enter the index database,
and adjust the order according to w.
End.

Figure 1 Ontology-based correlation algorithm flowchart theme

Experiment

I convened a number of long-term training in computer grade examination of computer science teachers. Based on their years of teaching experience, access to a variety of computer grade examination review books and materials, we can discuss the initial construction of the computer grade exam review ontology.

According to the National Computer Rank Examination questions, after discussion contains multiple-choice questions to determine the review, operating questions review, simulation tests and a series of related topics NCRE core vocabulary. In the experiment, the title weight T and body weight B were set to 0.6 and 0.4 respectively. The value of a djustment parameter K was set to 0.9. I searched the article 30, regarding the choice NCRE review in Baidu library where as the test set, and calculate the topic related to degrees.

Algorithm evaluation criteria can take Web information retrieval evaluation. Using precision(P), recall rate(R), significance level(F), ranking in the results page(OD).

$P = D/A$,D is actually collected the number of documents belonging to the topic. A is collected all of the number of documents.

$R = D/T$,D is actually collected the number of documents belonging to the topic. T is Test set actual number of documents belonging to the topic.

$F = 2 * (R * P) / (R + P)$.

Users have always liked more in line with their search intent sort the search results pages, so close to the user search intent of the page Od smaller the value, the more high ranking, users of the search algorithm is the higher degree of recognition.

In the experiment, the author designed a test set actual number of documents belonging to search for topics of 27. Namely,T=27. Based on the above proposed theme-based ontology similarity algorithm to determine the theme of the test set, and according to page Ontology-based correlation algorithm (denoted in the table Method 1) and user-based relevance feedback adjustment algorithm (Table noted in the method 2), compared to results retrieved. I first asked a student to speak their search intent, enter keywords in the 1st of search elements results shown in Table 1, the 10th to enter the same keywords, search elements results shown in Table 2.

Parameter	Algorithm in the paper	Method 1	Method 2
D	23	23	18
A	30	30	30
T-D	4	4	2
R%	76.6	76.6	60
P%	85.1	85.1	66
F%	80.6	80.6	63.1
Od	4	4	4

Table 1 The results of the first experiment compared

Parameter	Algorithm in the paper	Method 1	Method 2
D	25	23	18
A	30	30	30
T-D	2	4	2
R%	83.3	76.6	60
P%	92.5	85.1	66
F%	87.7	80.6	63.1
Od	2	4	2

Table 2 The results of the 10th experiment compared

From the data in the table 1, we can see that the user first input keywords, the search results of the algorithm described in this article and method 1 are almost the same. The precision and recall rate of Method 2 is slightly lower a bit, but as user searches and the user clicks on the trips increased, the concerned page rankings showed an upward tendency, The precision and recall rate of Method 2 changed little. The precision and recall rate of the algorithm described in this article showed an upward tendency. We can see that this algorithm is superior to Method 1 and Method 2.

Conclusion and future work

To take advantage of ontology thematic analysis, we can build on ontology based on semantic analysis to achieve, thus, the initial realization of human-computer interaction semantics, making the computer a clear message needs to improve topic discovery process accuracy. At present, the relevant articles on how to build theme body gradually increased, but there is not a good theoretical support, and ontology concepts are generally extracted by hand, systematic ontology faced a field. This makes large-scale ontology-based applications cannot carry out. Thus, in the above work to further improve on the basis of the ontology and how to use user feedback, establish user feedback model to get a better judge the accuracy of the theme will be the next focus of the work.

References

[1]Yi J，Sundaresan N.Metadata based web mining for relevanc[C]∥Proceedings of the In IEEE 2000 Inernational Database Engineering and Applications Symosium (IDEAS’00) Yokohama：IEEE,2010：113-121．

[2]Latifur Khan, Dennis Mcleod, Eduard Hovy. Retrieval effectiveness of an on tology based model for information selection[J], The VLDB Journal The International Journal on Very Large Data Bases, 2004, 13(1): 71- 85

[3]Yuri A Tijerino , Reza Sanati. Onto TEMAS: an ontology-based teaching materials search engine[J] . Journal of Computing Sciences in Colleges, 2005, 20(4):177- 182.

[4] Sergey Brin. Extracting Patterns and relations[J] Computer Science Department Stanford University，2009:172-183.

[5] Min zheng. Subject correlation degree decision algorithm based on dynamic matching. Microelectronics and computer In Chinese. 2009.

[6]Junfeng Song,Weiming Zhang. Research on information retrieval model based on Ontology[J]. Journal of Nanjing University: Natural Science Edition In Chinese. 2005, 41(2) :189- 197.

[7]Dezhi Xu,Weili Guo. Research on the algorithm of topic correlation based on Ontology. Journal of Yunnan University (NATURAL SCIENCE EDITION) In Chinese , 2007, 29 (S1) : 51-54.

[8]Xuefang He,Changhe Li,Zhenghao,Shi. Correlation degree adjustment algorithm based on user feedback. Journal of Shandong Agricultural University (NATURAL SCIENCE EDITION) In Chinese.2007, 38 (4): 615- 618.