Keywords Semantic Extension in Semantic Search Model

Xuejun Yu，Jing Lv

School of Software Engineering, Beijing University of Technology, Beijing, China, 100124
yuxuejun@bjut.edu.cn, lvjing0514@sina.com

Abstract - The goal of semantic search is to abandon the search model that based on simple keyword matching, combined with semantic technologies, make full use of the characteristics of ontology-based semantic information relationship to provide users with more high-quality and efficient search results. The study of this paper focuses on the semantic extension of keywords in semantic search process, to consider the vocabulary expansion with different factors, thus providing a set of factors that affecting the expansion of the vocabulary and the reasonable weight for these factors that verified by experiment.

Index Terms - ontology, semantic technologies, semantic search, keywords extension, weight.

1. Introduction

Semantic search model is built on the basis of the traditional search model, basic module processes consistent with the traditional search model[1]. The basic process is: First, crawl a large number of data resources from the Internet as the basic content of search, and then, classify and index the data which is collected. Finally, complete the design of the user interface so that users could get the content from the index library through the interface.

The goal of this study is added the semantic technology to the process above, built a semantic search model which is distinction and superior to the traditional search model, to achieve this goal there are three key issues to be addressed:

A. Preprocessing of the user query conditions

The main task is to extract useful information from the user's input, meanwhile, filter out the words which doesn't make sense to the retrieve results. This is the first step in semantic search, plays an important role in ensuring the accuracy of the preprocessing.[2]

B. Semantic expansion for the extracted keywords

This is the linking process between user query conditions and the built ontology concept library, because domain ontology is established by a large number of experts, had been studied, verified and expanded. It is an authoritative knowledge base, so matching the input information of users with the authority concept and extended in a reasonable way is very important to search results. Through that process, will get a set of terms which are more precisely described, and it will play an active role in the search results.

C. Complete the sort based on semantic relations from result

To measure a search engine just by the recall rate is one-sided, also need to focus on how to recommend the most accurate and the resources most likely close to the user' retrieval condition to the users. So it has high requirements for the sorting of search results. The sorting algorithm of semantic retrieval added semantic information to the original retrieval model, it could achieve the goal of recommending the key content by priority.

In the above three aspects, the key parts are part2 and part3. They are the key of establishing the relationship between the semantic information and retrieval resources, they will directly affect the results of semantic retrieval. This paper will be based on the second point to research.

2. The Methods of Keywords Semantic Relevance Extensiond

Traditional search engine is completely based on a set of keywords to retrieve, however, this rigidly keyword matching method would bring a series of problems. For example: the input vocabularies may be inconsistent with the terms used in the document, and the contents based on text matching is not the same as the user's retrieval intention[3]. While add the semantic information in keywords extension method will abandon the original vocabulary matching, give full play to the relationship about vocabularies in authority ontology, introducing the characteristics of the ontology can be reasoning. It will achieve the goal of effectively improve the retrieval results, especially in the recall ratio.

In this paper the calculation method of vocabulary expansion is divided into two categories, one is non-semantic vocabulary extension, the other is based on semantic information. The former is based on the characteristics of the Chinese from the vocabulary itself to analyze the relationship between keywords. The latter is the key point of vocabulary extension method, based on ontology, to consider from different factors, they are: hierarchical relationship, the connectivity, matching rate of instances and matching rate of properties. The specific methods are described below:

A. Non-semantic vocabulary extension

1) Part of speech and literal similarity: The part of speech plays an important role in languages. When two words' part of speech are different, extended set of vocabularies related to the initial meaning of the terms has undergone great changes. So judge the unity of the parts of speech between terms is a prerequisite[4-5].

In case the two words in the same part of speech, it need for the further calculation of literal similarity. It refers to the proportion of the same Chinese characters in words. In Chinese each character has its own specific meaning, therefore, in the same condition of speech, if the greater the proportion of the same Chinese character between the words, to some extent, means higher correlation. Summed literal similarity calculation method as (1):

$$wordSim(x,y) = size(x \cap y)/(sizex + sizey) . \quad (1)$$

$size(x \cap y)$ represents the number of the same Chinese characters between x and y. The denominator $sizex + sizey$ represents the value of the sum of Chinese characters.

2) The length of similarity: In Chinese, the length of the word has influence on the meaning. Because of each Chinese character has its own meaning, length of two terms the more close their meaning is often more similar. Calculated as (2):

$$sizeSim(x,y) = 1 - |sizex - sizey|/(sizex + sizey) . \quad (2)$$

$sizex$ represents the number of Chinese characters in x, and $sizey$ means the number of Chinese characters in y.

B. Semantic vocabulary extension

1) Factors of hierarchical relationship: When considering similarity based on ontology, hierarchical relation-ship is one of the important factors in an ontology net-work, a reference based on a fully-matched concept in the know-ledge network, the concepts at the same level or at the adjacent level is often having greater semantic correlation, more conducive to the expansion of knowledge. However, with the increase of the different between levels, the relevance of vocabularies is weakened, the necessity of expansion between the concepts gradually reduced. So take the hierarchical relationships of concept in authoritative ontology as one of the important factors to measure whether there is a need to expand between the keywords, calculated as (3):

$$levelSim(x,y) = \begin{cases} 1, level(x) = level(y) \\ 1/[|level(x) - level(y)| + 1], level(x) \neq level(y) \end{cases} . \quad (3)$$

$level(x)$, $level(y)$ respectively represent the level of the two concepts in the ontology.

2) Connectivity factor: If there exists connectivity relations between the concepts in ontology network, it indicates they have more or less relevance in that field, the relevance can be calculated. If there is no connectivity relations between concepts, that is to say they cannot be reached each other, in terms of there is no relevance in connectivity relations. Connectivity factors calculated as formula(4):

$$\text{connection}Sim(x,y) = \begin{cases} 1, \text{Equivalent concepts} \\ 0.8^n, 0 < n < 4 \\ 0, \text{Other situations} \end{cases} . \quad (4)$$

ConnectionSim (x, y) represents the keywords similarity in the connectivity factors. The first case in (4) is, ontology defines the two concepts completely equivalent, so connecting factors made this a maximum of 1; N in second case represents concept y is reachable from concept x through n routes. Here requires 0<n<4, i.e. the threshold of distance is 3. Only consider three situations which the reachable connectivity path length are 1,2,3, the longer distances are not considered. The relationship between weight of connectivity and distance is exponential relationship, the longer the distance will lead to lower weight, otherwise higher weight. The others in third case refers to there is no connectivity relationship between two concepts or the distance of path is greater than the threshold value 3. Under these circumstances, the correlation is so low that we can ignore them. So at this moment, the weight is 0.

3) Matching degree of instances and properties: Ontology, not only includes concepts, terminology in the field, but also contains the properties, and instances of the concept. The more same properties, the more similar to the way they used to describe the two concepts.

The matching degree of the instance is similar to properties. For example, the concepts of computer and laptop, instances of them will contain a large number of different brands and models of computers, the instances matching degree of this two concepts is much more higher than the concepts between computer and mobile phone. Therefore, the similarity of the instances and properties can be used as one of the important basis of whether the keyword is worthwhile to expand.

According to the principle described above, the calculating method of instances matching degree is as (5):

$$instanceSim(x,y) = num(instx \cap insty)/[num(instx) + num(insty)] . \quad (5)$$

The calculation method of properties matching degree is as (6):

$$propertySim(x,y) = num(propx \cap propy)/[num(propx) + num(propy)] . \quad (6)$$

3. The Confirmation of Keyword Semantic Relevance Extension Parameters

Discussion of above, analyzed factors of keyword expansion and the method of calculation from semantic to non-semantic. They would be the basis of measuring the key words when they are combined. The keywords semantic relevance extension could be calculated as (7):

$$\begin{aligned} relevance(x,y) &= k_1 \cdot wordSim(x,y) + \\ &k_2 \cdot sizeSim(x,y) + k_3 \cdot levelSim(x,y) + \\ &k_4 \cdot connectionSim(x,y) + \\ &k_5 \cdot instanceSim(x,y) + k_6 \cdot propertySim(x,y) \end{aligned} . \quad (7)$$

Relevance(x, y) has summarized the factors of keywords semantic extension and the method of calculation, however, each part of the parameters need to be further determined. The confirmation of these parameters, need to examine the actual

impact of each factor on the correlation effect. We combine the daily experience with experiment in the determined process. Because when the semantic search is proposed, its goal is to make the search engines can model on the judgment of human logic. So comparing the model calculation result with the empirical values is the most effective way. In (7), $k_1+k_2+k_3+k_4+k_5+k_6$ =1. Due to the key issue in semantic search is semantic extension, so we increase the ratio of semantic parameter, and decrease the ratio of k_1 and k_2.

The specific experimental design ideas and steps of parameters are: Select 20 keywords in the ontology of e-commerce as experimental objects. They are divided into 10 groups, each group has two words. According to the research results, assign weights to six parameters. Then substitute parameters into (7) and to calculate the similarity of the 10 groups words according to (7). The inaccuracy between the result of similarity and the daily experience could reflect the gap between semantic extension and the actual logic under the sequence of this set of parameters. The next step is to calculate the average inaccuracy value of 10 experiments, which will serve as the inaccuracy that set by the parameters for this group. Based on these 10 groups of experimental subjects, repeatedly adjust the distribution of the parameters sequence and calculate the average inaccuracy value respectively. The parameter sequence which has smallest inaccuracy as the final result set of the parameters. Table I is calculated under a certain set of parameters setting.

Mentioned above, the determination of whether the setting of parameters are reasonable is constantly adjusting the ratio of $k_1 \sim k_6$. Calculating the average inaccuracy value of these keywords in different conditions. Among them, when the average inaccuracy of parameters (δ) is minimum, take the group of parameters as the optimal similarity calculation parameters result sequence.

The formula of calculating the average inaccuracy is (8):

$$\overline{\delta} = (\delta_1 + \delta_2 + ... + \delta_n)/n = 1/n \cdot \sum_{n=1}^{n} \delta_n . \quad (8)$$

The formula of calculating the optimal average inaccuracy is (9):

$$\overline{\delta}_{\text{optima}} = \min\{\overline{\delta}_1, \overline{\delta}_2, ..., \overline{\delta}_n\} . \quad (9)$$

Through repeated experiment, the experimental results can be obtained: $\overline{\delta}_{optima}$=0.048. This indicates that in variety of circumstances, the minimum average inaccuracy is 0.048. At this moment, k_1=0.15、k_2=0.05、k_3= 0.1、k_4= 0.4、k_5= 0.1、k_6= 0.2. So using this set of parameters as the parameters results to calculate keywords semantic relevance extension values. Eventually, the formula is (10):

$$\begin{aligned} relecance(x,y) &= 0.15 wordSim(x,y) + \\ &0.05 sizeSim(x,y) + 0.1 levelSim(x,y) + \\ &0.4 connectionSim(x,y) + \\ &0.1 ins\tan ceSim(x,y) + 0.2 propertySim(x,y) . \end{aligned} \quad (10)$$

TABLE I A certain set of parameters setting

Parameter settings	Keyword x	Keyword y	Similarity results based on this parameter settings	Similarity results based on experience	Inaccuracy (δ)
k_1 =0.15	laptop	notebook	1.00	1.00	0.00
	computer	mobile phone	0.51	0.50	0.01
k_2 =0.05	cellphone	mobile phone	1.00	1.00	0.00
	computer	displayer	0.46	0.40	0.06
	notebook	Ultrabook	0.74	0.85	0.11
k_3 =0.1	mouse	keyboard	0.53	0.50	0.03
	mobile HDD	USB flash disk	0.62	0.70	0.08
k_4 =0.4	graphics card	mouse pad	0.40	0.40	0.00
k_5 =0.1	router	networking products	0.60	0.70	0.10
k_6 =0.2	RAM	bluetooth headset	0.34	0.25	0.09
	average inaccuracy: $\overline{\delta}$ =0.048				

4. The Use of Keywords Semantic Relevance Extension Methods

The research results presented in this study which is based on the semantic and non-semantic expansion methods to calculate the similarity of keywords. The formula as (11):

$$relevance(x,y): \begin{cases} \text{Based on non-semantic extension:} \\ relevance_1(x,y) = \\ k_1 \cdot wordSim(x,y) + k_2 \cdot sizeSim(x,y); \\ \\ \text{Based on semantic extension: } relevance_2(x,y) = \\ k_3 \cdot levelSim(x,y) + k_4 \cdot connectionSim(x,y) \\ + k_5 \cdot ins\tan ceSim(x,y) + k_6 \cdot propertySim(x,y) \end{cases} \quad (11)$$

Restore the formula for the semantic and non-semantic categories in order to facilitate the extension work in practice. After giving the calculation method, the next step is to consider how to apply the calculation method into actual keyword expansion, thus more fully reflect the advantages of semantic retrieval. Specific flowchart about the semantic relevance extensions for keywords is shown in Fig. 1.

Fig. 1 has described the process of semantic relevance extension for keywords, the two situations are:

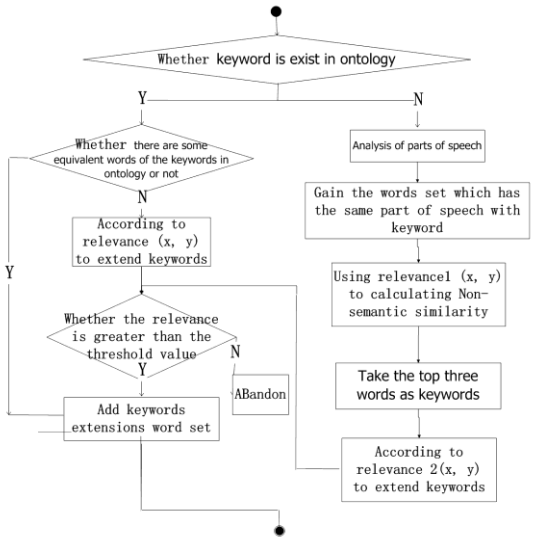

Fig. 1 Flowchart of semantic relevance extension for keywords

The query input by the user after pretreatment, when the filtered keywords exist in the current ontology. First of all, is to determine whether there is a vocabulary in the ontology which is equivalent of the keyword. If there is equivalent vocabulary, added to the collection of semantic extension keywords results directly; If the keyword does not yet exist equivalent words, it needs to measure the vocabularies in ontology and keywords comprehensively by relevance(x,y) to get the weights of similarity. Then filter out the vocabularies that have greater weight than the threshold from the vocabulary sets which have weight of correlation. The correlation threshold is set in advance in the system, the experiments proved that two vocabularies can be considered as a certain relevance when the correlation of vocabularies is greater than 0.6. Therefore, the correlation threshold is set to 0.6 in this paper.

The query input by the user after pretreatment, when the filtered keywords do not exist in the current ontology. At this time, the query words cannot be associated with ontology which has been built, it means there is no semantic relations between them. In this case, has to use non-semantic method to obtain the words which are most similar to the keywords input by the user. Then complete semantic extension for these words.

This approach is based on the analysis of parts of speech, gain the words set which has the same part of speech with keyword. Then use relevance1 (x, y) to calculating non-semantic similarity. To sort the results, we choose three results which have highest similarity as keywords. Because the non-semantic factors has already been considered, therefore, only to consider the semantic relevance in the process of similarity extension. According to relevance 2(x, y) to calculate, the subsequent process is similar to the first case. Screen out the words whose weight are greater than the similarity threshold as the final keyword expansion set. The vocabulary expansion has been completed.

5. Conclusion

By combining the needs of the semantic search model, proposed solutions to semantic extension of keywords -- the key issues of model in semantic and non-semantic ways. Identified six factors for vocabulary expansion, they are: part of speech and literal similarity, the length of similarity, hierarchical relations, connectivity factor, matching degree of instances, matching degree of properties. According to the actual impact to the result, allocate the weight for each factor. And the experiment proved the rationality of these weights. At last, given the usage of the semantic extension in actual searching process. The results of this study, expanded semantic vocabularies will in place of the original user input information as the actual conditions of semantic retrieval. This is a critical step in establishing the semantic information of retrieval model.

References

[1] Gong Cheng, Weiyi Ge, Yuzhong Qu. Falcons: Searching and Browsing Entities on the Semantic Web, International World Wide Web Conference, Proceeding of the 17th international conference on World Wide Web Beijing [M].China POSTER SESSION: Posters，2008:1101-1102 .

[2] Osma Suominen, Kim Viljanen, and Eero Hyvonen. Usercentric Faceted Search for Semantic Portals[M]. The Semantic Web: Research and Applications,2007:356-370.

[3] Zhang Mingbao, Ma Jing, Shi Xiuli. Research on field ontology applications in information retrieval[J]. Journal of The China Society for Scientific and Technical Information, 2010,29(2):215-222

[4] LiuLing DAI, Bin LIU, Yuning XIA. Measuring Semantic Similarity between Words Using HowNet[C]. International Conference on Computer Science and Information Technology 2008,2008.

[5] Jorg-Uwe Kietz, Raphael Volz, Alexander Maedche , Extracting a Domain-Specific Ontology from a Corporate Intranet[C]. Proceedings of the Fourth Conference on Computational Natural Language Learning and of the Second Learning Language in Logic Workshop, Lisbon, 2000.

[6] Wang S，Tanaka Y.Topic－oriented Query Expansion for WebSearch［C］.Proc of the 15th Int Conf on World Wide Web. New York: ACM，2006: 1029－1030