The optimization of campus information semantic retrieval ordering algorithm based on ontology

Xu ZHANG [a*], Dian CHEN [b], Jingru YANG [c], Zhao CHEN [d]

School of Information Science and Technology ,Beijing Forestry University,Beijing,Beijing,100083,China
[a]sxxyzx1994@163.com,[b]cddiandian@qq.com,[c]bbdejingru@qq.com,[d]cz71@bjfu.edu.cn
*15600608330

Keywords：Campus information; ontology; semantic retrieval algorithm; weigh; order

Abstract： In order to achieve a fast and efficient retrieval of the campus information as well as a better user experience, this paper has constructed an ontology model based on campus, realized and optimized the semantic retrieval algorithm. Meanwhile, by calculating semantic similarity and semantic relevancy, and according to the hit rate of campus information and browsing history of users, the model makes a real-time distribution and adjustment of weights of search keywords to make the sorting of retrieved results more close to the attention and interest of users so as to meet the needs of college teachers and students when they do information retrieval.

Introduction

With the rapid development of Internet, the resource of campus information has become increasingly diverse. For more and more teachers and students, the demand of gaining the campus hot spots of information is also increasing. The traditional web retrieval based on keyword couldn't meet people's demands any more. Conversely, the semantic retrieval technology based on ontology can make semantic analysis and contextual analysis and obtain correlative retrieval to respond user demands and get higher precision ratio.

However, the ordering only depending on semantic retrieval can’t please users for most of retrieval results. And the simply semantic retrieval is so difficult to determine the weight of different keywords for the search term including a few keywords that the actual ordering can’t satisfy users need. In order to understand the user intent more clearly and improve the recall ratio and precision ratio, this paper proposed that on the basis of semantic retrieval based on domain ontology combining with the user clicks is to adjust the weight of search term and change the ordering.

Campus is a tiny life system which includes teaching, entertainment, food and shopping. Because of the complexity and variations of campus life, the need of campus information presents the characteristics of diversity. But the information technology which can achieve diversified information service is still deficient and the development of campus information is a bit slow now. So building the campus domain ontology based on campus review platform, taking the advantage of domain ontology in semantic relation and reasoning mechanism can form a new college campus high-quality service and efficient retrieval model.

The ontology and the campus domain ontology

Ontology is a clear formal specification for the shared conceptual model[1] which includes four meanings: the concept model（Conceptualization）, definitude（Explicit）, formalization（Formal）and sharing（Share）. Generally speaking, ontology is composed of conceptual class, relations, functions, axiom and examples. And the ontology make the conceptual semantic description by forming the classification and the relationship between the concept. Actually, the key role of ontology is to build the domain model.

The domain ontology is a concept model that is used for describe the specific domain. It aims to catch the knowledge and provide a common understanding of knowledge about the field. At the same time, it can give the clear definition of relationship between words from different levels of formal schema[3].

The ontology model based on campus used the seven steps[4] to establish campus resource ontology and it's constructed by protégé which is the editor of ontology. The campus ontology consists of four small ontologies which are about classes, food, recreation and shopping in this paper. In theory we should create corresponding core classes for each of ontology and determine the attributes and relationships of concept. On this basis, we should describe explicitly concept of different features and properties as well as constraints and fill the individuals [5]. With food ontology as an example, we constructed the ontology based on above ideas and follow the steps below.

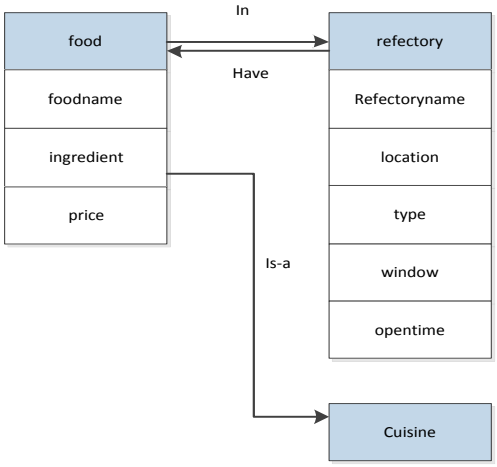

Chart 1 the relationship of classes in food ontology

The first step is to establish the core classes. In food ontology, food and refectory are the core classes. And the data properties of food include flavor, ingredients, price and so on. The data properties of refectory include positions, types, windows and open time.The second step is to supplement[6] the class structure. Then determining the object properties[7] of concepts namely the relationship between classes is required. The relationship between the classes on food ontology is showed in chart 1.The relationships shown in the picture correspond to the object properties of ontology. Learning from the chart 1, food is a kind of cuisine in the refectory. In the two relationships, food is the domain, but refectory and cuisine are the range of relationship. The fourth step is to determine the data properties[8]. In food ontology, food is the domain of dish name, ingredient, price and comment. Similarly, refectory is the domain of locations, the names of refectory, types, windows and open time.

The design of retrieval sorting process

On the basis of building the campus domain ontology, in order to meet the needs of user group, optimizing the semantic retrieval sorting algorithm is necessary. Users enter a search term through the search box at first, the system can get the key words set by using the technology of words segmentation[9] and obtain the initial sort based on the similarity and correlation algorithm. If the click rate of searching result changed, the system would adjust the weight of key words depending on the new click rate, otherwise keep the current weight. After completing the sorting, system will adjust the proportion between similar words in accordance with the relevant indicators. The process is shown in chart 2.

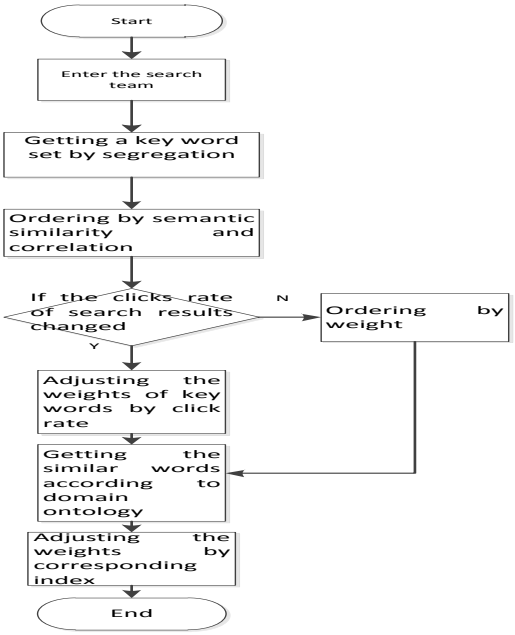

Chart 2 the process of retrieval

The similarity and correlation algorithm in semantic retrieval

In the semantic retrieval system based on domain ontology, the system computes the concept of similarity and correlation based on the ontology for user query expansion. The computing of concept similarity based on ontology mainly considers three elements as follows: semantic distance, semantic overlap ratio and node layer[10].

Computing the similarity with three elements:

$$Sim(x,y) = \frac{a}{Distance(x,y)+a} \times b \times \frac{|NodeSet(x) \cap NodeSet(y)|}{|NodeSet(x) \cup NodeSet(y)|} \times \frac{1}{r \times |Level(x) - Level(y)| + 1} \quad (1)$$

*The semantic distance of x and y is represented by Distance(x,y).*The node set from n nodes to root nodes is represented by NodeSet(n). *The level of node n is represented Level(n).*

The original semantic distance is represented by alphabet a when similarity is 0.5. The effectof adjusting semantic overlap ratio for similarity is represented by alphabet b. And the effect of adjusting level difference for similarity is represented by r.

Computing the correlation:

$$Rel(x, y) = \frac{l}{ShortestPath(x,y)+l} \quad (2)$$

Shortest(x,y) is the shortest of nodes x and y.The shortest of concepts is represented by l when correlation is 0.5

The relationship of similarity and correlation:

$$Sim\_Rel(x,y)=Sim(x,y)+Rel(x,y)-Sim(x,y)\times Rel(x,y) \quad (3)$$

The comprehensive index of x and y is represented by Sim_Rel(x,y).

The algorithm of adjusting key words weight based on user click behavior

The conditions of constructing algorithm Although, the search results by the method of similarity and correlation algorithm based on ontology are comprehensive and the method has higher recall and precision than general retrieval's,every key word of search term must have different proportion in the user view and the proportion don't be known by algorithm. The computer can't analyze users' love, but this paper proposed we can deduce the user need by user behavior and compute the weight of key words for adjusting the ordering and satisfying the user need.In order to adjust the key words of search term by user click, the system need to get the number of click by statistical method. According to obtain massive click of search term and corresponding result, the system can satisfy the public need.

First of all, *a set* $Q=\{q_1,q_2 \ldots q_j \ldots q_n\}$ *that* consists of key words which are segregated from user search term is formed. And the initial search results are shown by using the similarity and correlation algorithm based on ontology and ordering. The system can get the key words which correspond with the key words of set Q from every search result to compose a set $A_i=\{a_{i1},a_{i2} \ldots a_{ij} \ldots a_{in}\}$. a set $C_j=\{c_{1j},c_{2j} \ldots c_{ij} \ldots c_{mj}\}$ which is the group of click rate from m(number) results including the j(number) key words and similar words reflects the tendency of user click and user need.

The implement process of algorithm On the basis of obtaining the tendency of user click, the system adjusts the weight of key words. Because the result is based on the semantic retrieval, the weight of jth(number) key word is not only influenced by the corresponding key word, but also influenced by the similar corresponding key word. And the influence coefficient is the combination of similarity and correlation. Under their mutual influence, the system obtains the weight of jth key word in all key word of the search term.

$$Rqj = \frac{c1j}{\sum_{i=1}^{n} cij} \times r1j + \frac{c2j}{\sum_{i=1}^{n} cij} \times r2j + \cdots\cdots \frac{cmj}{\sum_{i=1}^{n} cij} \times$$

$$rmj = \sum_{i=1}^{m} \left(\frac{cij}{\sum_{i=1}^{m} cij} * rij\right) \quad (4)$$

The comprehensive value between j key word and the similar words of m result is represented by $R_j=(r_{1j},r_{2j}, \ldots\ldots r_{mj})$ $(j=1\sim n)$.As the number of search key words is tremendous, if the system needs to compute the weight according to the above algorithm every time, the burden of computer would be bigger and bigger and the speed of search would be slower, even impair the user experience.As a result, based on the semantic retrieval, the system uses the similarity and correlation between the key words and similar words to construct the proportion of key words and other words.The proportion of j'key word and kth word which is similar with j key word.

$$Rqjk = \frac{Rqj * rjk}{Rqj'} \quad (5)$$

Through the above algorithms, the weight of key words has been confirmed and the system has obtained the satisfactory search result for users.

Experimental analysis

This paper investigated 100 students who study at Beijing Forestry University aiming at 200 records about campus food in database based on similarity and correlation algorithm and recorded the satisfaction and recall rate of retrieving. Then, this paper did the same experience after using the new algorithm at one week. Compared with two results, the two experimental data are shown in the following table.

Table 1the data of experimental results

The query method	satisfaction	Recall ratio
Semantic retrieval	73.48%	87.64%
The combine of semantic retrieval and weights adjusting	84.93%	85.96%

From the above results by comparison, the new algorithm adjust the weight of key words and improve the satisfaction and recall rate to meet public need.

Conclusion

On the basis of similarity and correlation algorithm, this paper proposed an algorithm that can adjust the weight of key words based on user click rate. Compared with current similarity and correlation algorithm, the optimized algorithm segregates the key words and satisfies user need by analyzing the user behavior.

Acknowledgements

Project sources: Beijng forestry university"Beijing university students' scientific search and Entrepreneurship plan" NO.S201410022075

References

[1] Studer R, Benjamins V R, Fensel D. Knowledge Engineering, Principles and Methods[J]. Data and Knowledge Engineering,1998, 25(1/2): 161-197.

[2] J LIANG, Q LIU, A YU. Personalization recommendation algorithm for Web resources based on ontology[J]. Journal of Computer Applications, 2014,34(11):3135-3139. (In Chinese)

[3] Z DENG, S TANG, M ZHANG.Overview of ontology[J].Acta Scicentiarum Naturalum Universitis Pekinesis, 2002, 38 (5) : 730-733. (In Chinese)

[4] J GAN, Y JIANG, Y XIA.Ontology and its application[M]. Science Press,2011:98-102. (In Chinese)

[5] G ZHANG, Y CHEN, Q ZHOU et al. Information semantic relativity retrieval based on domain ontology[J].Computer Engineering, 2011,37（20）：34. (In Chinese)

[6] J GUO.The building of domain ontologies and its application[D].Beijing:School of Computer Science,Beijing University of Posts and Telecommunications,2007,32-33. (In Chinese)

[7] Z FENG，W LI，X LI. The semantic project and applyment [M].Beijing：Tsinghua university press, 2007:67-69.(In Chinese)

[8] H LIU, D XU. The summarize of semantic similarity and correlation computing based on ontology[J]. the science of computer,2012,39(2):8-13(In Chinese)

[9] X SHAO.Chinese word segmentation technology reserch based on lucene[D].Xian: University of Electronic Science and Technology of China，2012,15-19. (In Chinese)

[10] H LV, W SONG, R YANG.Weighted semantic similarity algorithm based on domain ontology[J].Computer Engineering and Design，2013,34（12）： 4210-4213. (In Chinese)