Web Collection Of Key Technique Based On Domain-Oriented Ontology

SONG Mei Mei
Department of Information and Engineering
Shandong First Medical University & Shandong
Academy of Medical Sciences
Ji Nan, China
mmsong.shuxue@163.com

SA RongBo
School of Life Sciences
Shandong First Medical University & Shandong
Academy of Medical Sciences
Ji Nan, China
sarongbo@163.com

***Abstract*— Proposed a new key technology to solve web collection algorithm. Firstly, the domain-oriented ontology is established, Computing domain-oriented ontology correlation, Secondly, it ascertains the relevance of appraisal links and web pages of domain-oriented ontology. This correlation algorithm considers that the number of keywords is more. The ability to show the theme is stronger. Lastly, it decides whether the goal of content is related with pages.**

Keywords—search engine Web collection domain-oriented ontology the precision rate the recall rate (key words)

With the development of the internet and more and more attention of information resources， we have more and more relied on internet to gain the information we need. Network has come to pervade every aspect of our lives. The personalized search engine has developed for more than ten years. We have made gratifying achievements in the research of synchronization of network information and user's actual needs. Search engine how to better learn user interest, Service aspects that provide different information content to different users using different service strategies and methods is not enough to meet people's needs, The limitation of the this lack of personalized search engine is becoming more and more prominent, In the information search from generalization to personalized and intelligent direction development, web information acquisition plays a pivotal role. But the speed of information collection is more and more unable to meet the actual needs. And the traditional web collection is to collect information pages as much as possible. According to the survey, even large information collection systems have only 30-40 percent coverage of resources. This process reflects the speed and quantity of the collection, not much attention to the order of the collection and the relevant topics of the collected pages. One of the great advantages of doing this is that it is relatively simple to achieve. However, this traditional collection method can’t update the rapidly increasing information resources in time, It is unable to improve the pertinence of personalized information retrieval to user services. Therefore, the research of web page information collection technology based on user personalization is of great value and significance.

Establishment of domain-oriented ontology

Subject information collection requires that the object searched be as consistent as possible with the target in the process of searching for resources. Generally speaking, this similarity based calculation can’t reflect the semantic information of web documents very accurately. Ontology is a new knowledge organization tool used to formalize the concepts and terms of a specific domain. It provides the basis for the standardized description of resources. Domain-oriented ontology represents the concept of a discipline domain and the relationship between concepts. After the establishment of domain-oriented ontology, through the methods of concept implication, attribute association, mutual constraint and axiom, the domain-oriented ontology is organized into a formal ontology model with network structure, which can be shared. Not structured information is transformed into information entity with a certain structure, which becomes the information that the machine can understand, and provides support for semantic interoperation and intelligent reasoning. This means is the development direction of information retrieval language. More and more researches apply ontology to the field of information retrieval.

Constructing domain-oriented ontology

Domain-oriented ontology can abstract a domain into concepts and relationships between concepts, and express conceptual relationships as conceptual semantics. Formal coding process of Domain-oriented Ontology using corresponding Ontology description language, Domain-oriented ontology represents conceptual semantics such as objects, relationships, and classes in a single word collection. The class structure of domain-oriented ontology refers to the description of the relationship between classes, which is generally divided into upper and lower relationships, global local relations, synonyms and so on.

Implementation of domain-oriented ontology building algorithm: The OWL ontology is processed by Jena API, and the OWL file generated by domain-oriented ontology library is parsed and inferred by Jena tool, which can improve the semantic matching ability of the system. The model uses OWL language rule, OWL as a network ontology language. It can formalize domain-oriented ontology. It can describe the relationship between classes and attributes by axioms and constraints, and has strong semantic and knowledge representation ability to achieve a unified understanding between the machine and the user.

Domain-oriented Ontology correlation Analysis algorithm

The attributes and hierarchical relationships of concepts play an important role in describing the semantics of concepts. The method of keyword correlation analysis is used to calculate the relevance of this topic, which is essentially the number of given keywords appearing in the statistical page. This correlation algorithm believes that the more keywords appear, the stronger the ability to represent the topic. The real problem, however, is that a semantic concept and keywords can be a one-to-many or a many-to-one relationship. Therefore, the correlation analysis method based on keywords is not very efficient in analyzing these special cases. If there are too many pages related to keywords, the dimensionality should be reduced before calculation, and a large amount of computer resources will be consumed. Based on the calculation of semantic similarity of concepts and the comprehensive use of ontology hierarchy and conceptual attributes to calculate the comprehensive similarity between words and expressions, it is more suitable for domain ontology to analyze the subject relevance based on concepts. This method takes into account various relationships between concepts, starting with page analysis, replacing the same concept keywords with the theme words with concepts, and calculating the similarity between concepts in the domain. The page and theme feature vector can be used to judge the correlation at semantic level and reduce the dimension of computing vector.

Topic correlation analysis algorithm:

Input: domain-oriented ontology , Preprocessed web pages , Regulating parameter

Output: Conclusions on whether the page is relevant

The algorithm includes the following steps:

(1) Replace each keyword v_i and keyword v_i' with the same concept with concept k_i (The upper meaning of the m level of v_i) Semantic page p_j to p_j' .

(2)The content of a web page is of different importance when expressing the subject. For example, the title usually describes the main content of the page, and bold italics may be the key content of the page. Think of the content of a web page as a tree-like arrangement, each of the same html tags as a concept, Use C to represent a new set of concepts in the web content, After resource integration, the concept set is made up of n components. Remember $C=(C_1,C_2,...,C_n)$,The eigenvector of the resource vector Xc(d)=(C_1w(d1);,…,C_iw(di);,…, Cnw(dn)) In which w(di) is the concept of domain-oriented ontology, C_iw(di) is the weight (frequency) of the domain-oriented ontology in the C_i text. For example, the theme computer is not an isolated concept, it is located in a figure 1 below

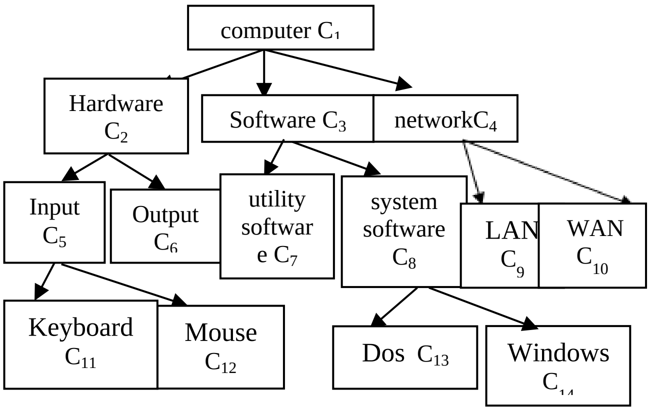

Fig. 1 A case of a ontology tree

Many of the concepts in the parent-son relationship are related to topic C1, but to varying degrees, the similarity between concepts is defined by the following formula [2]:

$$sim(w(di),w(dj))=\begin{cases}\dfrac{\alpha\times(di+dj)}{(L(w(di),w(dj))+\alpha)\times 2\times MaxL\times max(|di-dj|,1)} & di\neq dj\\ 1 & di=dj\end{cases}\quad(1)$$

$sim(w(di),w(dj))$ represents the similarity between two concepts, di,dj represents the level in the tree, $L(w(di),w(dj))$ represents the shortest path between two concepts, $MaxL$ represents the maximum depth of the tree, $\alpha\geq 0$ is an adjustable parameter. Finally, the correlation between C and the subject is calculated.

$$sim(Xc(d),w(d1))=\sum_{w(di)\in c\cap w}c_i w(di)sim(w(di),w(d1))\quad(2)$$

$sim(Xc(d),w(d1))$ represents the similarity between Cand a topic, $sim(w(di),w(d1))$ represents the relevance of each domain-oriented ontology concept to the topic.

（3）Assuming that the threshold is λ , when $sim(Xc(d),w(d1))$ is greater than λ , the web page is related to the topic, and when $sim(Xc(d),w(d1))$ is less than or equal to λ , the task is independent of the web page.

Experiment and analysis

1.1 Collection based on page analysis and evaluation

Step one: according to the topic, select a certain number of terms in the professional field as candidate feature words, and build the dictionary.

step two : selecting the webpage related to the theme as the training text , calculating the frequency of the candidate words in the webpage , and then calculating the weight of each candidate word in the texts , for example , the weight of the i characteristic word is expressed as Wi , and the weight is obtained by the TF-IDF formula [3] .

Step 3: each training text is represented as an n-dimensional feature vector U={w1,w2,…,wn}, so that the training text is mapped to a point in vector space, and the average weight of each candidate feature word in these training texts is obtained by arithmetic averaging method. If the weight value is large, the value is chosen, and the user candidate feature is mapped to a point in the vector space, which constitutes the subject feature vector V=(v1,v2⋯vn). The information matching in document space is transformed into vector matching, and the topic similarity formula is obtained.

$$Sim(U,V)=\cos(U,V)=\frac{\sum_{i=1}^{n} wi*Vi}{\sqrt{\sum_{i=1}^{n} wi^2 * \sum_{i=1}^{n} Vi^2}}$$

Step 4: given a threshold λ, if the correlation degree is greater than λ, it is considered to be related to the topic. λ is half of the average value of the similarity between each training text and the topic. Experimental data as shown in Table 1

1.2 omain-oriented Ontology based on Collection

Step 1: semantic tagging of each web page based on the created ontology, and use of web crawler to collect the page.

Step 2: the content of the web page has different importance in expressing the topic. Taking the ontology tree shown in figure 1 as an example, the similarity between the concepts is calculated.

$$sim(C_{11},C_7)=\frac{5\alpha}{6(5+\alpha)}, \quad sim(C_{11},C_5)=\frac{5\alpha}{6(1+\alpha)}$$

It can be seen that the greater the distance between the concepts, the lower the similarity.

Step 3: assuming the threshold is λ, λ takes the average value of the calculated similarity. When $sim(Xc(d),w(d1))$ is greater than λ, the web page is considered to be related to the topic, and when $sim(Xc(d),w(d1))$ is less than or equal to λ, the task is independent of the web page.

According to the above experimental steps, the proposed domain-oriented ontology correlation analysis is compared with the topic correlation method of vector space model based on keywords. The experiment is to download 150 pages from the webpage containing computer subject by two kinds of topic correlation analysis method, extract topic and digest as text, divide the text, form feature set, extract feature vector, analyze the correlation degree, and compare the analysis results (the keyword weight is set as 1, m =3, related page threshold. λ =10); 20 categories of computer professional words, such as hardware, software, input, output, application software, etc., are selected, and m level feature words. The results of the experiment are shown in Table 1.

Table 1 Performance analysis and comparison of correlation

Method	Processing the number of pages / pages	Actual related web pages / pages	Time consuming	Web pages larger than threshold	Actual collection of related pages / pages	Precision %	Recall %
Correlation analysis based on key words	150	120	307	94	88	93.62	73.33
Correlation degree analysis based on domain-oriented Ontology	150	120	267	114	103	90.12	85.83

It can be seen from Table 1 that the accuracy and recall rate of the topic-dependent method based on ontology is significantly higher than that of the topic-related method based on keywords. The main reason is that the semantic extension method of ontology concept is adopted in the topic-related method based on ontology. The similarity can also be calculated for web pages with no obvious interest in keywords in some web pages, which is superior to the keyword calculation method in this respect. Moreover, the time consumed by the method proposed in this paper is also reduced accordingly.

Information collection plays an important role in the accuracy of information retrieval. Based on the peer research, this paper proposes a domain ontology based information collection method, and analyzes the domain correlation of the information collection results. The experimental results show that this method is more effective than using keyword vector model. The accuracy of the relevant pages, recall rates have increased significantly.

Main references:

[1] HUSTON G. Interconnection, peering and settlements—Part I [J].Internet Protocol Journa,l 1999, 45(3): 124-136.

[2] CHEN Jie，JIANG Zu Hua，Conceptual similarity calculation of domain ontology ，Computer engineering and Application，2006，33:163-165

[3] Lewis D D et al. Training algorithms for linear text classifiers[C] . In Proceedings of the Nineteenth International ACM SIGIR Conference on Research and Development in Information Retrieval. New York : ACM Press,1996.298-3