Concepts Discrimination Research

JIANBO XU[1], CHENGWEI MA[2], JIAXUN CHEN[3]

1,3: Information College

Donghua University

Yan'an Road 1882, Shanghai, 200051

PEOPLE'S REPUBLIC OF CHINA

2: Aviation University of Air Force

PEOPLE'S REPUBLIC OF CHINA

Abstract: Concepts discrimination is proposed to improve precision for information retrieval. Based on Ontology, semantic correlation between the ontology concepts and the user keywords of information retrieval is measured to fix on a semantic context that the computer can understand. Concepts synonymous extension is used to improve the probability of matching the concepts with the user keywords. Several coefficients definitions are used to denote the correlated extent of ontology concepts and the user keywords. Sorting these coefficients is to choose the most suitable concepts for the user keywords. With these suitable concepts , A distinct semantic context for the user keywords is worked out.

Keywords: Ontology, Concepts Discrimination, Semantic Correlation

1 Introduction

Information retrieval needs high precision and high recall factor. Conventional processing method is to retrieve with keywords, only documents including these keywords will be retrieved. Despite some documents' semantic information satisfying the user's requirement, they will not be retrieved. The recall factor is debased. In another condition, some documents include one or two key words of the all specified keywords. But the documents don't satisfy the semantic information of the users' requirement, because many words may have several meanings. The precision will be debased[1]. The core problem is that the retrieval system can't recognize the concerns of users correctly only depending on the keywords retrieval, so it is necessary to recognize the semantic information which is expressed by those keywords by some efficient processing methods.

In order to improve the recall factor, semantic annotation based on Ontology is used to describe the information source[2]. In order to improve the precision, we use semantic distinction based on Ontology to analyze these keywords to avoid the mismatching concepts. Concepts' semantic discrimination based on Ontology is the topic of this paper. Note that, the arithmetic of concepts matching in this paper is based on the characters matching of Chinese word items, so the word items of concepts are all expressed in Chinese and explained in English.

2 How to raise precision

In the concept distinction processing based on Ontology,

the key step is to complete the suitable mapping between the ontology concepts and the user keywords. Through mapping to the concepts of the ontology, the concepts contained in these keywords are discovered. In the mapping process, we need to assure that the irrelevant concepts will not be mapped and associated, and on the other side correlative concepts will not be omitted[3].

In fact, when a set of keywords combined together are presented to retrieval engine, these keywords fix on a related semantic context. While the computer doesn't have enough comprehension ability to choose the correct semantic context. In this paper, Ontology technology is introduced to enable computer to measure the distances and correlations between the concepts included in the keywords and the concepts in the Ontology, then the best suitable concepts will be gained to express the users' intention by purging the irrelative concepts.

Ontology is a formal, explicit specification of a shared conceptualization[4]. According to the different targets of defining the ontology concepts, Ontology can be classed to 4 kinds of Ontologies, including top-level Ontology, domain Ontology, task Ontology and application Ontology[5]. In this paper, some research about domain Ontology are described.

In the primary definition of the ontology concepts, a concept is expressed with one word item in the ontology system, and sketch map are showed on fig. 1. In fact, every word in our life has many synonyms. When a user specifies a word to retrieve information, he/she may not use the word item listed in the ontology concepts system, and perhaps another synonymous word is used as keyword. In this condition, the specified keywords can't be efficiently matched, and the ability of the ontology system catches the intention of the user is degraded. In order to resolve this problem, concept's word list in ontology system is extended with synonymous words[6] .

3 How to extend synonyms for ontology concepts

In the process of researching on how to use Ontology in information retrieval, we set up an experimental ontology concepts system. In this ontology system, the topic of "Environmental protection" is chosen as domain to organize concepts. The ontology system is described as directed acyclic graph, shown as fig.1. Every node in the map is a concept, and every concept is expressed with on word item which is unique in the whole ontology concepts system. Now, according to synonym thesaurus and specialty resources, we do some synonymous extension for these ontology concepts[7][8]. Every concept in the ontology concepts system consists of a primary word item and several secondary word items. The primary word item is the unique word item in the old basic concepts system. In the following description, we obey the rules: the set of word items including the primary word item and the secondary word items is called as *ontology concept synonyms list*, every word item in the list is called as *ontology word items*, and a specified keyword is called as *user keyword*, and several specified keywords are called as *combination of user keywords*.

For example, for the concept of "固体废弃物"("solid offal") in the ontology concepts system with synonymous extension, "固体废弃物"(solid offal) is the primary word item, while "固体废物"("solid waste") and "废渣"("rubbish") are the secondary word items. These three word items composed the ontology synonyms list ("固体废弃物"("solid offal"), "固体废物"("solid waste") and "废渣"("rubbish")) for the concept of "固体废弃物"(solid offal), shown as fig.2.

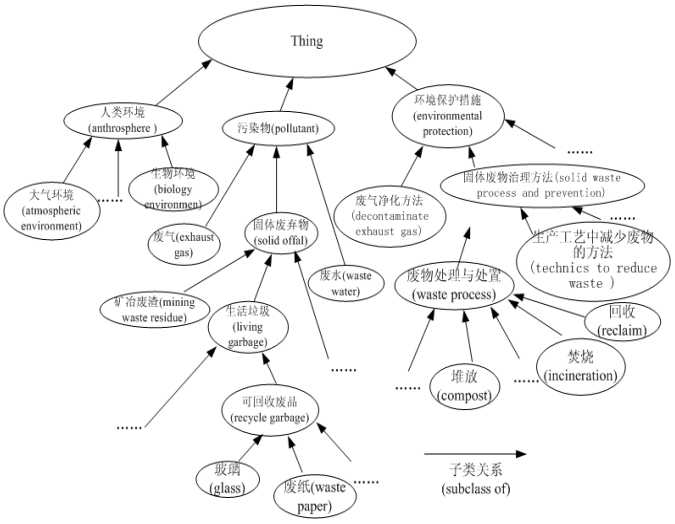

Fig.1. Sketch map of ontology concepts system of environmental protection

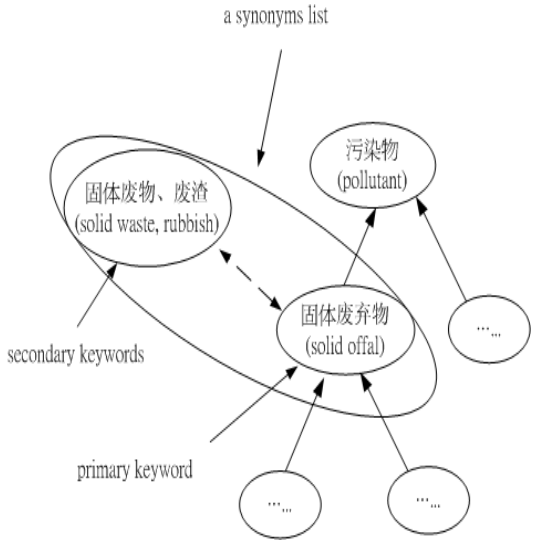

Fig.2 Sketch map of synonymous extension for ontology concepts

When the matching concept is retrieved for a user keyword, the user keyword is compared with every item in the *ontology concept synonyms list*. In this process, the probability of successfully matching the user keyword with the ontology concept is greatly improved. Of course, in this condition, a user keyword may match one or more ontology concepts, and perhaps fully matching, or partly matching, so semantic relation between the user keyword and the concepts must be considered. We use some algorithms to select the most suitable ontology concept to express the meanings included in the user keywords.

4 Concepts Discrimination

Through retrieving the ontology synonymous lists in the ontology concepts system with DFS (Depth First Search) or BFS(Breadth First Search), we match the user keywords with some ontology concepts[6]. Every chosen concept is called as *query concept*. These query concepts may have several different meanings. For a user keyword, there may be several concepts matching with it, so there are mismatching concepts. *The target is to remove these mismatching concepts and choose the most suitable concept for the user keyword*. The correlation among these concepts should be decided based on semantic distance. We define a coefficient as a token of their correlation.

Every concept has an *ontology concept synonyms list,* and the word items in the list are marked as WI_1, WI_2, … WI_i, …, WI_n, abbreviated as $\{WI_i\}$. The word items are compared with user keywords one by one, then the matching word items are figured out. Based on the number of matching characters of the word item (WI_i) which match with the user keyword, a coefficient is defined for the word item(WI_i) , called as WI_coeff. The biggest coefficient of all the word items in the *ontology concept synonyms list* is defined as a coefficient of the query concept, called as con_coeff. On the other side, when there is correlation between of two query concepts, a coefficient called as correlated_coeff is defined. In the following paragraph, definition 1,2 and 4 are proposed by us, and definition 3 is cited.

Definition 1. Word item coefficient of a query concept: For a word item WI_i in the *ontology concept synonyms list* of the query concept, the coefficient WI_coeff_i is a token of the matching degree between the word item WI_i and the user keyword KW_j.

WI_coeff_i = the matching characters'number of WI_i

and KW_j /(the number of the characters of WI_i + the number of characters of KW_j - the matching characters'number of WI_i and KW_j) ---------- (1)

Example 1: The word item "废纸"(waste paper) and the user keyword "纸张"(paper) are compared, and the matching character is "纸"(paper). The number of matching character is 1, and the number of the word item and the user keyword all are 2, the number of all the other character in the word item and the user keyword is 3 (that is the result of "2+2-1"), so WI_coeff = 0.333. While another word item "废弃纸张"(abandoned paper) is compared with the user keyword "纸张"(paper), then the coefficient is 0.5.

Definition 2. Coefficient of query concept: For a query, the coefficient is a token of the matching degree between the query concept and the user keyword. A bigger coefficient shows closer relationship between the query concept and the user keyword. The number of items in the *ontology concept synonyms list* is n.

$$Con\_coeff = \max WI\_coeff_i \quad 1 \le i \le n \text{ ---------- (2)}$$

Example 2. the word item "废纸"(waste paper) and the other item "废弃纸张"("abandoned paper") all are the word items of the concept "废纸"(waste paper), so the bigger value of the two word items' coefficient is chosen, and Con_coeff of the query concept for the user keyword is 0.5.

Definition 3. Semantic distance is defined as $SD(QC_i, QC_j)$ [9]: The semantic distance between two query concept QC_i and QC_j is the shortest distance between the two concepts in the ontology concepts system. The ontology concepts system is organized as directed acyclic graph. According to the characteristic of directed graph, if the distance between two adjacent and linked nodes is 1, the shortest distance between any two nodes can be figured out based on the arithmetic of Dijkstra[10]. This definition is used to measure the correlation between two concepts. Smaller semantic distance, more correlated between the concepts.

Definition 4. Correlation coefficient of query concepts: For a set of user keywords （KW_1, … , KW_i,…, KW_m）, several set of corresponding query concepts can be fixed on. For every user keyword KW_i, there is a query concept set, marked as $QCSet_i(i \in [1,m])$. Because the number of query concepts matching with every user keyword is different, the number of the query concepts corresponding with KW_i is marked as Len_i, the query concept set $QCSet_i$ can be expressed as $\{QC_{i1}, QC_{i2}, \ldots, QC_{ij}, \ldots, QC_{iLen\_i}\}$.

$$Correlated\_coeff_{ij} = Con\_coeff_{ij} + \sum_{k=1}^{i-1} \frac{Con\_coeff_{k.TOP}}{SD(QC_{ij}, QC_{k.TOP})} + \sum_{k=i+1}^{m} \sum_{r=1}^{Len\_k} \frac{Con\_coeff_{kr}}{SD(QC_{ij}, QC_{kr})}$$

$$i \in [1,m], j \in [1, Len\_i], k \in [1,m] \text{ and } k \neq i, r \in [1, Len\_k]$$

---------- (3)

$$Ratio_{kr} = \frac{Con\_coeff_{kr}}{SD(QC_{ij}, QC_{kr})} \text{ ---------- (4)}$$

$$Sum_k = \sum_{r=1}^{Len\_k} ratio_{kr} = \sum_{r=1}^{Len\_k} \frac{Con\_coeff_{kr}}{SD(QC_{ij}, QC_{kr})}$$

---------- (5)

The meanings of the definition:

(1) For the query concept QC_{ij} , Con_coeff_{ij} is a token of the matching degree between QC_{ij} and the user keyword KW_i.

(2) The second part and the third part of Equation 3 are the token of the correlation degree of the query concept QC_{ij} ($i \in [1,m]$, $j \in [1,Len\_i]$) with other user keywords KW_k ($k \in [1,m]$ and $k \neq i$) whose corresponding query concept set is $QCSet_k$, Len_k is the number of the query concepts in $QCSet_k$. The $Ratio_{kr}$ is a token of the matching degree of QC_{kr} 、KW_k and the correlation degree of QC_{kr} 、QC_{ij}. Before the most suitable concept for the user keyword KW_k is fixed on in $QCSet_k$, the Sum_k is used as a token of the correlation degree of QC_{ij} and KW_k ($k \in [i+1,m]$); As the most

suitable concept marked as $QC_{k.TOP}$ is fixed on and the coefficient of the chosen concept is marked as $Con\_coeff_{k.TOP}$, the ratio of $Con\_coeff_{k.TOP}$ and $SD(QC_{ij}, QC_{k.TOP})$ is a token of the correlation degree of QC_{ij} and KW_k ($k \in [1,i]$).

(3) For a set of $Correlated\_coeff_{ij}$（$j \in [1,Len\_i]$）, the corresponding query concept which has the biggest $Correlated\_coeff_{ij}$ is considered as the most suitable concept for the user keyword KW_i .

(4) In a general way, the number of user keywords are equal to or less than 3. In this condition, the computing complex is not too high.

Example 3. the user keywords is the combination of "纸张"、"回收"("paper", "reclaim") . The user keyword "纸张" ("paper") is marked as KW_1, and The user keyword "回收" ("reclaim") is marked as KW_2. The query concepts chosen for the user keyword KW_1 are {"废纸" "废弃纸张"}("waste paper", "abandoned paper"), marked as QC_{11}, and {"白纸"} ("white paper"), marked as QC_{12} ,while the query concepts set chosen for the user keyword KW_2 are {"收集"、"回收"}("collect", "reclaim"), marked as QC_{21} , and {"召回"、"回收"}("recall on", "reclaim") , marked as QC_{22}. According to the equation 2, Con_coeff of query concept QC_{11} is 0.5, Con_coeff of query concept QC_{12} is 0.333, Con_coeff of query concept QC_{21} is 1, and Con_coeff of query concept QC_{22} is 1.

The semantic distance between the query concepts QC_{11} and QC_{21} is 9 , and the semantic distance between the query concepts QC_{11} and QC_{22} is 15. So, for the query concept QC_{11},

Correlated_coeff=0.5+1/9+1/15=0.678.

The semantic distance between the query concepts QC_{12} and QC_{21} is 11, and the semantic distance between the query concepts QC_{12} and QC_{22} is 17. So, for the query concept QC_{12},

Correlated_coeff=0.333+1/11+1/17=0.483.

The two Correlated_coeff are compared, and the query concept QC_{11} is reserved as the more suitable ontology concept for the user keyword KW_1.

After the ontology concept corresponding to the user keyword KW_1 is fixed on, for the query concept QC_{21} , Correlated_coeff=1+1/9=1.111； for the query concept QC_{22} , Correlated_coeff=1+1/11=1.091。 So, the query concept QC_{21} is reserved as the more suitable ontology concept for the user keyword KW_2.

5 Arithmetic description

According to the definitions and equations, the arithmetic is defined as follows:

Step 1: the number of the user keywords is marked as m, and the user keywords is KW_1, KW_2, …, KW_i,…, KW_m ($i \in [1,m]$).

Step 2: For a user keyword KW_i, retrieving in the ontology concepts system with DFS(Depth First Search) or BFS (Breadth First Search), we will get a query concept set $QCSet_i$($i \in [1,m]$). The number of query concepts in the set is marked as Len_i, then the set $QCSet_i$($i \in [1,m]$) is expressed as { QC_{i1}, QC_{i2}, …, QC_{ij}, …, QC_{iLen_i} }($i \in [1,m], j \in [1,Len\_i]$). According to Equation 2 , we figure out Con_coeff_{ij} for the query concept QC_{ij} .

Step 3: IF m = 1 then

QC_{11}, …, QC_{1j}, …, QC_{1Len_1} are sorted in descending order of Con_coeff_{1j} , the sequence is listed below:

QC_{1o}, QC_{1p}, QC_{1q}, ……(o,p,q∈[1,Len_1])

And $Con\_coeff_{1o} \geq Con\_coeff_{1p} \geq Con\_coeff_{1q} \geq$……

So QC_{1o} responding to Con_coeff_{1o} is reserved as the most suitable ontology concept for the user keyword KW_1, and other query concepts are given up.

Step 4: Else /* m > 1. */

Step 5: For each $QCSet_i$($i \in [1,m]$)

Step 6: For each QC_{ij} ($j \in [1, Len\_i]$)

According to Equation 3 , $Correlated\_coeff_{ij}$ is figured

out.

End of For each QC_{ij}

Step 7: QC_{i1}, QC_{i2}, ……, QC_{ij}, ……, QC_{iLen_i} are sorted in descending order of $Correlated\_coeff_{ij}$,the sequence is listed below:

QC_{io}, QC_{ip}, QC_{iq}, ……$(o,p,q \in [1,Len\_i])$

And $Correlated\_coeff_{io} \geq Correlated\_coeff_{ip} \geq Correlated\_coeff_{iq} \geq$……

So QC_{io} responding to Con_coeff_{1o} is reserved as the most suitable ontology concept for the user keyword KW_i, and other query concepts are given up.

Step 8: End of For each each $QCSet_i$ (in Step 5)

Step 9：End of IF (in Step 3)

6 Conclusion

When keywords are used to retrieve information, the recall factor and precision are low. For this situation, ontology annotation technology is used to describe the information in website or information database to improve the recall factor. And in this paper, a concept distinction technology based on Ontology is used to improve precision. Actually, in the information retrieval experiment the precision of the retrieval is improved. But the response is delayed because of more computing needed. This solution will be optimized in the future research.

Reference

[1] Sujian Li,Houfeng Wang,Shiwen Yu and Chengsheng Xin. “News-Oriented Automatic Chinese Keyword Indexing”. In Proceedings of The Second SIGHAN Workshop, on Chinese Language Processing .ACL2003 :92-97

[2] T. Athanasiadis, V. Tzouvaras, K. Petridis, F. Precioso, Y. Avrithis and I. Kompatsiaris: “Using a Multimedia Ontology Infrastructure for Semantic Annotation of Multimedia Content”, 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2005) at the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, Nov. 2005.

[3] Lu Zhimao, Liu Ting, Li Sheng ,"Unsupervised Chinese Word Sense Disambiguation Based on Equivalent Pseudowords" ,Proceedings of the International Conference on Chinese Computing 2005 , ICCC2005

[4] Gruber T R. A Translation Approach to Portable Ontology Specification. Knowledge Acquisition, 1993, 5:199~220.

[5] Guarino N. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In: Pazienza M T, eds. Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, Spring Verlag, 1997, P 139~170.

[6] Sabrina Tiun, Rosni Abdullah, Tang Enya Kong, "Automatic Topic Identification Using Ontology Hierarchy",Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing ,2001,Pages: 444 - 453

[7] Dong shubao, 《Process and Reuse the Solid Waste》(2^{nd}) , Publish House of Metallurgy Industry，1999.

[8] Mei Jiaju, Zhu yiming,《Synonymous thesaurus》Shanghai Publish House of Thesaurus，1983.

[9] Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. In: IEEE Transactions on Systems, Man and Cybernetics. (1989) 17—30.

[10] Yan Weiming, Wu Weiming, 《Data Structure》: language C, Beijin : Tsinghua Publish House ,1996, P168, 189.