# Research on Web Cross Language Information Retrieval Based on Domain Ontology

Xiaorong Cheng, Haojun Guo, Yuhui Wang, Wei He
*School of Computer Science and Technology,*
*North China Electric Power University, Baoding, HeBei, 071003, China*
*haojunguo@126.com*

## Abstract

*With the increasingly rich of Web resources, people need to cross-language information and knowledge sharing, Cross-Language Information Retrieval(CLIR) research on such problems. This paper brings forward an Web cross language information retrieval model based on the domain ontology. The model basing on the technologies of traditional information retrieval, use domain ontology to describe the relevant domain knowledge in different kinds of languages, comprehend and extend query terms, lead the retrieval rise to semantic level. The main ideal and methods of the model are described in detail, and through experiment to validate our approach. The experiment are designed to retrieval travel news in Chinese from Sina website with query in English ,its results prove that this model can improved the average precision/recall of retrieval to certain extent.*

## 1. Introduction

Cross Language Information Retrieval is a method that query by one language and retrieved others language information. Where query language known as the source language, the language of retrieved object known as target language. With the rapid expansion of Web information resources and constantly enrich of multilingual information, CLIR has become a key factor of global knowledge-sharing.

At present, CLIR research is active in European languages, meanwhile, achieved more satisfactory results[5].However, due to Chinese have no spacing between words and with other self characters, how to realize the translation between Chinese and other foreign languages is still a difficult problem. Therefore, CLIR research on Chinese and other languages is relatively less. Currently, English and Chinese cross language retrieval have following main limitation: ①source language and target language translation is inaccurate; ②retrieval is often word-matching rather than meaning matching, average recall / precision rate is low.

The emergence of Ontology and Semantic Web technology can improve this status. Over the past decades, ontology is used to the computer field for the knowledge sharing, integration and reuse. Ontology as "a formal specification of a conceptualization"[2],it comprise a standard terminology glossary as well as some semantic description of the norms. Semantic Web is an extension of the current World Wide Web, can be divided into three layers, basic assertion layer, schema layer and logic layer[2].These information have well-defined semantics, it is able to do well in the computer and human cooperation. The emergence of them, made the information on the Web were given a good definition of meaning, it can be understand and process for machines, which possibly makes to process information automatically on the Web, the comprehensiveness and accuracy of retrieval will be greatly enhanced. This paper will combine the present English-Chinese translation and semantic-based related information retrieval technologies to design and realize a Web English-Chinese cross language information retrieval system based on domain ontology.

## 2. Key technologies of model

### 2.1. Query translation based on bilingual domain ontology

At present, cross language information retrieval system mostly use query translation mode, queries will be translated into the language used by the documents, because it is relatively simple to realize. Language translation method mainly include machine-based, machine-readable dictionary, text corpus-based[1].The machine translation is usually use a complete sentence, but in information retrieval, the query sentence often comprises some query keywords. Dictionary translation is a simple method, but it is hard to dealing with the words that have more than one meaning.

In the cross language information retrieval process, the translation effects have a direct impact on the accuracy of follow-up retrieval results. For ordinary users, query sentence often comprises simple keywords, lack of

978-0-7695-3290-5/08 $25.00 © 2008 IEEE
DOI 10.1109/CCCM.2008.291

the integrity of semantic information, machine translation based and dictionary based translation methods can not achieved satisfactory results. The system according to above limitation, adopt bilingual domain ontology to solve the problems semantic loss and distortion when translating between query language and retrieval language.

Bilingual domain ontology is the concrete express form of domain ontology in bilingual languages, it is similar with the bilingual semantic dictionary. Norm of synonym in two different languages is a key feature of bilingual domain ontology, that is the meaning of corresponding concept from two language ontology are same. This paper, build English-Chinese bilingual domain ontology which introduced the norm of synonym, so the concept can mutual corresponds between English and Chinese. The common connotation of the concept can not use any words of a language to express, only to describe its meaning, use number or symbols to identify. For example, we can use "123" to identify the concept of "one who travels for pleasure", its corresponding English vocabulary is "tourist", Chinese counterpart is "旅行者".

Domain ontology-based CLIR used domain ontology as interlingua to regulate the concepts of bilingual domain ontology, made the expression is consistent of connote - tion between source language and target language, and established bilingual mapping based its meanings, to eliminate the gap between language expression and connotation of concept[3].The relation between ontology and bilingual ontology as shown in Figure 1.

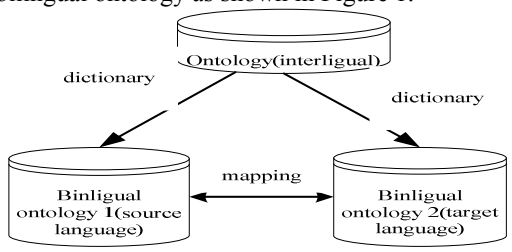


**Figure 1 Relation between ontology and bilingual ontology**

The paper established corresponding relationship between concept of bilingual domain ontology and every term of documents and query. Suppose query $q_E$= {$t_1$, $t_2$, $t_n$} is terms set of English for $q_E$ in order to mapping, defined function SE for every $t_i$ in $q_E$, record as $S_E(t)$ = {c ∈ C|t ∈ QE(c)}. So, query $q_E$ can be expressed with concept $S_E(t_i)$,record as $CS_E(t)$={ $S(t_1)$, $S(t_2)$,…,$S(t_n)$}.

For the keywords are not in bilingual domain ontology, use dictionary to translate. If some keywords still dissatisfy with above two step, these words will be OOV, record them and not be translated, these can be record in the translation log waitting for manager to translate by manual and add its corresponding meaning in the bilingual ontology later.

Since the system is based on specific domain to retrieval, for the words which are out of ontology, the deviation of translation results will not too great by use professional dictionary to translate and select first meaning as translation.

## 2.2. Concept similarity calculation

In the semantic similarity based information retrieval, the paper draws on "partial matching" strategy of the vector space model, compute its similarity as the basis to taking or rejecting document after getting the semantic vector of the user query and the document. Mainly consider two aspects when compute the similarity of concepts: the concept-correlation value in the document set and the hierarchy structure similarity in the ontology.

The calculation of concept-correlation value adopt the idea in paper[10],if the correlation degree of two concepts is higher, the context environment of their place should more like. Let $f_{ij}$ is the frequency of occurrence in the document corpus of concept $C_i$, these document corpuses include $C_j$, each concept uses these frequencies to constitute vector, $w_i = (f_{i1}, f_{i2}, ...f_{ij}, ..., f_{in})$ represent the vector of concept $C_i$. The correlation value between $C_i$ and $C_j$ can be computed as follows:

$$Sim_d(C_i, C_j) = \frac{w_i \bullet w_j}{|w_i| \bullet |w_j|} = \frac{\sum_{k=1}^{n} f_{ik} \bullet f_{jk}}{\sqrt{\sum_{k=1}^{n} (f_{ik})^2 \bullet \sum_{k=1}^{n} (f_{jk})^2}} \quad (1)$$

The hierarchy structure similarity computed as:

$$Sim_s(C_i, C_j) = \begin{cases} (1 - \frac{\alpha}{dep(R(C_i, C_j)) + 1}) \times \frac{\beta}{d(C_i, C_j)} \times \frac{son(C_j)}{son(C_i)}; & d(C_i, C_j) \neq 0 \\ 1; & d(C_i, C_j) = 0 \end{cases} \quad (2)$$

Where $dep(R(C_i,C_j))$ is the nearest root concept depth between $C_i$ and $C_j$, $d(C_i,C_j)$ is the distance between $C_i$ and $C_j$, son(C) is the number of subtree nodes for C in the ontology. $\alpha, \beta$ are tuned factors, we take $\alpha, \beta$ =0.5.

Combined (1)(2),for concept $C_i$ and $C_j$, the similarity is computed as:

$$Sim(C_i, C_j) = \delta \cdot Sim_d(c_i, c_j) + (1 - \delta) \cdot Sim_s(c_i, c_j)$$
$$\delta \in [0,1] \quad (3)$$

Where $\delta$ is tuned factors, we set $\delta$ =0.7.

## 2.3. Semantic query expansion

Users often use one or a few keywords to express their retrieval request, but it is sometimes difficult to express

their request loyally and accurately, this paper use domain ontology to understand user's retrieval request, and use its good concept hierarchy to some logical inference, then find some generalization and refining queries with initialization.

Definition1 (user query) the initial set of concepts what are user input[10].It can be expressed as:

$$Q = \{C_1, C_2, ..., C_i, ..., C_j, ..., C_m\}$$

$\forall i, j : 1 \le i, j \le m$ , where $C_i, C_j$ is user query keywords, and $C_i \neq C_j$

Definition2 (super class set)

$$\{F_i\} = \{C_i \mid \forall C_j, C_j \in subclassof\ (C_i)\}$$

Definition3 (sub class set)

$$\{S_i\} = \{C_j \mid \forall C_i, C_j \in subclassof\ (C_i)\}$$

Definition4 (Equivalence set)

$$\{E_i\} = \{C_j \mid \forall C_i \in F_i, C_j \in subclassof\ (C_i), C_i \neq C_j\}$$

For the user's query $Q_w\{(C_1, W_1), (C_2, W_2), ..., (C_m, W_m)\}$,where $W_i$ is the weight of $C_i$ for a query Q, the initial and subclass concepts weight W=1,the super class and equivalence concepts weight can be computed by formula(3), $W=Sim(C_i,C_j)$.

## 2.4. Weight calculation

Let D be a Web document set, $D=\{d_1,d_2, \ldots d_n\}$,and $Q_c$ be the set of query concepts, $Q_c=\{C_1, C_2 \ldots C_m\}$, $\{C_i\}$ be the set of $C_i(1\le i\le m)$and its synonyms, $f_{ij}$ is the number of occurrences in $d_j$ （$1\le j\le n$） of the set $\{C_i\}$,by adapt of the TF-IDF algorithm, the weight of $\{C_i\}$ in $d_j$ computed by formula(4):

$$w_{i,j} = \log(f_{ij} + 1) \times \log(n_i / n + 1) \quad (4)$$

Where n is the number of Web documents, $n_i$ is the number of documents attached to $\{C_i\}$.

Similarly, $\forall C_i \in Q_c$ , the weights of $\{F_i\}, \{S_i\}, \{E_i\}$ and its synonyms in D are computed by an adaptation of formula(4),record as respectively. The final weight $w_{id}$ is computed as:

$$w_{id} = w_{i,j} \times k_1 + w([F_i], D) \times k_2 + w([S_i], D) \times k_3 + w([E_i], D) \times k_4$$
$$k_1 + k_2 + k_3 + k_4 = 1 \quad (5)$$

Where $w_{id}$ is the average weight of Ci and its expansions in D, $k_1,k_2,k_3,k_4$ are tuned factors representing the relative "importance" of $\{C_i\}$, $[F_i]$, $[S_i]$, $[E_i]$.

For a given query $Q_c$, computed the weights $Q_c$ in each document, $d_j$= （$w_{1j},w_{2j},\ldots,w_{nj}$） ,where $w_{nj}$ is the weight of nth concepts in jth document. Similarity, computed the weight of every concept in $Q_c$, $Q_c$= ($w_{1c},w_{2c},\ldots, w_{nc}$), where $w_{nc}$ is the weight of nth concepts in $Q_c$. Now, the similarity measure between document d and the query $Q_c$ is computed as:

$$f = \sum_{i=1}^{m} w_{i,j} \bullet w_{i,c} / \left(\sqrt{\sum_{i=1}^{m} (w_{i,j})^2 \bullet \sum_{i=1}^{m} (w_{i,c})^2}\right) \quad (6)$$

## 2.5. Information retrieval algorithm

Because domain ontology knowledge is limited, so the query Q from translation module can be divided into two parts: one part is keywords that be included in ontology, record as { $Q_o$ } ; another is record as $\{Q_k\}$, the keywords in $\{Q_k\}$ are not in ontology[5].Retrieval algorithm are following:

```
Input: query Q
Output: RetrievalResults
Begin
  RetrievalResults={};
  TempResults={};
  If  Q is null
      Query end;
  Else
    Begin
    ①classify Q, divide it into two parts, record as { Qo} , {Qk} respectively;
    ② translate {Qo}based on bilingual domain ontology, translate {Qo}based on dictionary;
    End;
   run query function();
   return RetrievalResults;
End
```

The query function () as follows:

1)If $\{Q_o\}$is not null, concepts and relations in $\{Q_o\}$mapped for the concepts and relations in ontology, formed new concept set C;Else go to step (7);

2)Find synonyms for C, record as S,{C+S} be the query term($k_1$=1), inferencing by RDF inference model and retrieval, if the retrieval results are null, go to step(3),else return the retrieval result set, record as TempResults, go to step(6);

3) Find the direct subclass and its synonyms for {C+S}, record as SUB, if SUB is null, go to step (4), else SUB be the query term ($k_3$=1), inferencing and retrieval, return the retrieval set TempResults, go to step (6);

4) Find the direct super class and its synonyms for {C+S}, record as SUPER, if SUPER is null, go to step (5), else SUPER be the query term ($k_2$=1), inferencing and retrieval, return TempResults, go to step (6);

5)Find the equivalence set and its synonyms for {C+S} , record as SEMI, if SEMI is null, go to step(7),else SEMI be the query term ($k_4$=1), inferencing and retrieval, return TempResults, go to step (6);

6) If $\{Q_k\}$is not null, the keywords in $\{Q_k\}$and their

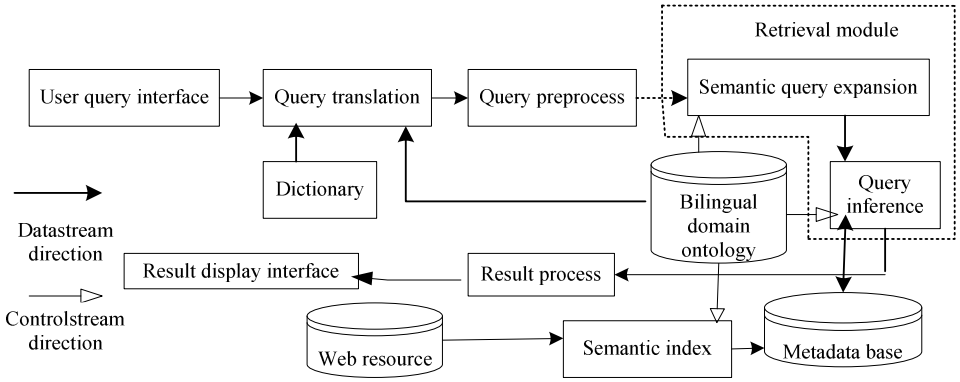


**Figure 2 English-Chinese information retrieval model based on domain ontology**

synonyms directly match with TempResults, return matching results, the results is the final results Retrieval Results, else TempResults is Retrieval Results.

7) Keywords in $\{Q_k\}$ are directly matched with document set; get the final retrieval results, record as RetrievalResults.

## 3. Domain ontology based web cross language information retrieval model

Based on the above analysis, this paper creates bilingual domain ontology to solve the problems semantic loss and distortion when translating between query language and retrieval language. Meanwhile, due to the keywords combination lack of semantic interpretation, recall rate and precision rate is low for keyword retrieval; we also use of the bilingual ontology to comprehend and expand user's intend. Design English-Chinese information retrieval model based on domain ontology as shown in Figure 2.

The model mainly have three parts, bilingual ontology, metadata base (index base) and retrieval module, the main functions are as follows:

[1] Bilingual domain ontology

Bilingual domain bilingual ontology in this model played an important role, it replaced dictionary to translate query terms. At the same time, mapping the user's query to the concept which in the domain ontology, and then through the link of concepts to query expansion.

[2] Index base

First, the model pretreated the Chinese document which will be retrieved, using the most positive word-matching method to segment sentences, remove stop words, other meaningless words such as empty word. In order to improve the efficiency of query matching, use the method of double-character-hash-indexing which in paper [11].Annotate and analyze the documents refer to bilingual domain ontology, transformed the concepts which in documents and include in ontology to terms of domain bilingual ontology. Finally, stored all the indexes into index-base.

[3] Retrieval module

Query terms matched with indexes recur to bilingual domain ontology, matching adopt vector space model, compute the similarity between vector of query and vector of document, sort in accordance with similarity, return to the users.

The workflow of the system are as follows: ① Establish domain ontology that related with information resources in specific areas, and the information in domain ontology encoded into RDF/XML format to provide information retrieval module[4]. ②Crawling specific Web resources, stored them in a Web resource base, instance which obtained by semantic index Web resource and stored in the semantic metadata-base. ③ Users input English query terms through user query interface, after translated by translation module, the translated query terms is delivered to the query preprocess module and transformed to the structured query terms, then submitted to the information retrieval module. ④Information retrieval module according to domain ontology knowledge, execute query and return the results, these results meet query request. ⑤ After processed by the result process module, the retrieved results are delivered to the display interface and displayed to the users.

## 4. Experiment and result analysis

The experiments use a bilingual travel domain ontology, which is described with RDF. The retrieval objects are 500 Chinese documents, all these from Sina Website and belong to travel field. Using keywords CLIR based on dictionary translation, Chinese information retrieval based on semantic and the retrieval of this paper proposed.

The value of Precision-Recall is response to a query. However, in fact is usually through the implementation of various queries to evaluate the retrieval algorithms. In this paper, use average recall/precision as a measure of system's performance indicator, precision(P) and recall(R) which is defined as follows:

$$P = \frac{R_a}{C}; R = \frac{R_a}{A}$$

Where C is the number of documents that meet the certain conditions of retrieval, A is the number of all relevant documents, $R_a$ is the number of retrieved documents.

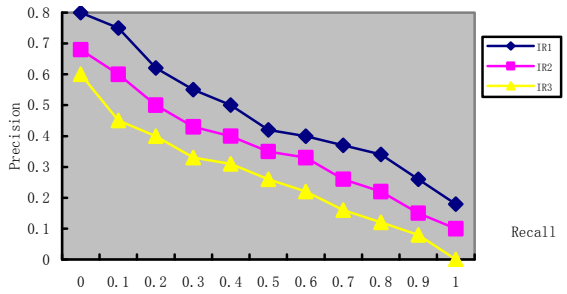


**Figure 3 Precision-recall**

**Table 1 Average precision of CLIR**

| IR System | Average Precision | % of $IR_1$ |
|---|---|---|
| $IR_1$ | 0.471 | 100% |
| $IR_2$ | 0.365 | 77.5% |
| $IR_3$ | 0.266 | 56.5% |

The average precision of the three IR model are reported in Table 1, and the precision-recall curves are shown in Figure3. Where $IR_1$ is the semantic retrieval use of Chinese language as query. $IR_2$ is the retrieval of this paper proposed method, first, input English terms and translate these into Chinese, then query information by use of section 2 methods.$IR_3$ is keywords CLIR based on dictionary.

When recall is near zero, the precision of IR1 nearly 80 percent, the precision of IR2 nearly 70 percent,IR3 is 60 percent, by analyzing documents, IR1's precision not reach 100 perfect, the main reason is that the domain ontology is not perfect, at the same time, the semantic annotation is also not accurate enough. IR2 retrieval use the ontology of IR1 and the information resources be semantic indexed, the performance of IR2 is not good than IR1 mainly due to the corpus of translation process is not big enough, some terms which out of corpus only rely on dictionary translation, at the same time, the OOV has not been translated at the first time. When recall nearly 1, the precision of semantic based retrieval higher than keywords based retrieval distinctly.

## 5. Conclusions

The ontology has very extensive application foregr-ound in the knowledge management, the Web information retrieval. In this paper, by virtue of CLIR and semantic retrieval, on the basis of the *Domain Ontology based Web Cross Language Information Retrieval* ,we present a new Web cross language information retrieval model based on bilingual domain ontology, the model's translation and retrieval strategy appropriate for cross language retrieval in special domain, to some extend, the problem of retrieval recall and precision low is better solved. Certainly, the realization of this paper proposed is still not perfect, construct bilingual ontology is difficult, put it in practice also need the development of Semantic Web technology and the idea which is put forward in this paper further improve. The next step work will introduce the ideal of interactive retrieval to improve the model's average precision/recall.

## References


[1] J.Y.Nie,M.Simard,p.Isabelle,etal.Cross-Language Inform--ation Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web[C],22thACM-SIGI R ,Berkeley,1999, 74-81.

[2] Berners-Lee T. Semantic web road map[EB/OL]. http:// www.w3.org/DesignIssues/Semantic. html, 2002-09-10.

[3] Volk M,Vintar S,Buitelaar P.Ontologies in cross-Language information retrieval.[2006-04-02].http://www.ling.si.se/ DaLi/volk/papers/WOW_Lucerne_2003.pdf

[4] Pablo Castells,Miriam Fernandez, David Vallet. An Adapta tion of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE TRANSACTIONS ON KNOELEDGE AND DATA ENGINEERING,VOL 19. NO. 2.FEBRARY 2007,261-271.

[5] Wang Jin,Chen Enhong,Zhang Zhenya.An Ontology-Based Cross Language Information Retrieval Model[J].Journal of Chinese Information Processing, vol.18 No.3,1-8.

[6] Wu Dan,Wang Huilin.The Mechanism of Ontology Appli-ed to Cross Language Information Retrieval[J]. LIBRARY AND INFORMATION SERVICE,vol 50, 2006, 10-13.

[7] Urvi Shah,Tim Finin,Anupam Joshi.Information Retrieval on the Semantic Web[C].Proceedings of the ACM Confer-ence on Information and Knowledge Management,2002.

[8] Min-Yuh Day,Chorng-Shyong Ong,and Wen-Lian Hsu. Question Classification in English-Chinese Cross Language Question Answering:An Intergrated Genetic Algorithm and Machine Learning Apporach.IEEE International Conference on Information Reuse and Integration, 2007.Aug. 2007:203-208.

[9] Wei Gao,Cheng Niu,Jian-Yun Nie.Cross Lingual Query Suggestion Using Query Logs of Different Languages[C].SIGIR'07,Amsterdam,The Netherlands.

[10] Song Erwei,Liu Zongtian.The Realization Mechanism of the Web Information Retrieval Based on the Domain Onto-logy.Computer Science,2007,34(5):104-110.

[11] LI Qing-hu,Chen Yu-jian,SUN Jia-guan.A New Dictionary Mechanism for Chinese Word Segmentation[J]..Journal of Chinese Information Processing, vol.17 No.4,13-18.