Journal of Applied Sciences

ISSN 1812-5654

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 13 (20): 4249-4260, 2013
ISSN 1812-5654 / DOI: 10.3923/jas.2013.4249.4260

An Approach of Ontology Construction and Semantic Query Expansion for High-speed Railway Domain Knowledge

[1]Ziyu Liu, [2]Xiaoming Zhang and [2]Xuehui Li
[1]School of Economics and Management,
[2]School of Information Science and Engineering,
Hebei University of Science and Technology, Shijiazhuang, Hebei, 050018, China

Abstract: Domain ontology can be used to extend the user's retrieval words and index domain knowledge. This study discusses the semantic retrieval system based on semantic query expansion using domain ontology constructed by thesaurus and thematic words for High-speed Railway Knowledge. Firstly, according to the actual situation of high-speed railway domain that is comprised of different professional fields, this study puts forward a construction methodology of domain ontology based on thesaurus and thematic words of high-speed railway that is built by the experts. Guided by this method, we build various majors ontology of high-speed railway and then they are preliminary merged into one unified domain ontology of high-speed railway. Secondly, this paper designs the model of semantic query expansion based on domain ontology and gave five kinds of semantic expansion methods. Then the model of semantic query expansion is applied in semantic retrieval system. Thirdly, we design a semantic retrieval system and the model of semantic query expansion is applied in it. Finally, this study makes a comparison between traditional keyword search method and our semantic retrieval method. The experimental results show that the developed semantic retrieval system can improve the recall and precision.

Key words: High-speed railway, thesaurus and thematic words, building ontology, semantic query expansion

INTRODUCTION

Ontology is playing an increasingly important role in the fields of software engineering, artificial intelligence, information retrieval and web service research (Klein and Bernstein, 2001). According to the level of dependence on research fields ontology can be divided into Top Ontology, Domain Ontology, Task Ontology and Application Ontology (Jin, 2001). Domain ontology can effectively organize the knowledge of that domain and make it easier to share and reuse.

High-speed railway (the new railway with speed above 250 km h^{-1}) is the development mainstream of China. The High-speed railway development strategy of China is "introduction, digestion, absorption and innovate", so we should draw lesson from forming advanced experience. In order to give facilities to query information about design, planning, construction and operation of high-speed railway, the building of High-speed Railway Fundamental Information Database System is important. High-speed Railway Fundamental Information Database System (Z2006-094) is an important science and technology project of Railway Ministry of China. Under the support of this project and through a variety of ways to gather information, the number in document base of literature that is related with high-speed railway is more than thirty thousands. Because "thesaurus of railway" is not suitable for high-speed railway, so under the support of the project we organize experts to rebuild "thesaurus and thematic words of high-speed railway" which is apply to high-speed railway literature information. Among these the first class category is ten, the second category is 64 and the third category is 208, 2488 subject headings Corresponds these categories.

At the same time we also build "foundation data table of high-speed railway". In order to search needed information, we need a good search pattern. Now using "thesaurus and thematic words of high-speed railway", we indexed the literature in document base and realized keyword-based search model. However, such a model misses the actual semantic information of the text and it leads to the low recall and precision. In order to deal with this issue, ontologies are proposed for knowledge representation which are nowadays the backbone of semantic web applications (Kara *et al*., 2012). Semantic

Corresponding Author: Ziyu Liu, School of Economics and Management, Hebei University of Science and Technology, Shijiazhuang, Hebei, 050018, China

query expansion based on ontology can improve efficiently the recall and precision. So we will use it in our new search system for high-speed railway knowledge.

To achieve semantic retrieval, it needs to build domain ontology firstly. Most domains, for example, agriculture, aviation, railway and high-speed railway, have their own vocabulary, such as dictionaries, thesauri and Thesaurus, so building the initial core ontology of industries domain based on the traditional classification/Thesaurus is a more scientific approach. Domain ontology can be used to extend the user's retrieval words and index domain knowledge. Under which we can achieve semantic retrieval for domain knowledge.

The semantic extension of user's retrieval words is also the more important step in semantic retrieval system. The designed system consists of four main components: Ontologies building, Retrieval words semantic extension, semantic precomputation for document and semantic retrieval. The experimental results show that the developed semantic retrieval system can improve the recall and precision.

RELATED WORKS

There are two major methods for the construction of domain ontology: One is from the perspective of knowledge engineering to research the construction method of ontology that is called ontology engineering, another is transform the existing vocabulary resources directly to ontology that is called thesaurus-based ontology construction method.

Query expansion is useful when there is a lack of correspondence in the terminology used in indexing and querying (Segura *et al*., 2011). When queries are correctly defined (i.e., are less ambiguous), there is more chance of success. The focus is therefore on poorly formulated, ambiguous queries or those containing specific terminology. Query expansion enables a search for resources with variations of the original query terms and even the addition of new terms resulting from the disambiguation (Bhogal *et al*., 2007). It is a query refinement method that aims to increase augment the level of relevance of the results obtained by adding new terms to an original query (Croch and Yong, 1992; Qiu and Frei, 1993).

The following are related works in ontology building method and semantic query expansion.

ONTOLOGY BUILDING METHOD

Ontology engineering: Emphasize to build ontology according to certain norms and standard is the main feature of ontology engineering. Up to date, the following is the well-known typical ontology and methodology that occur in the development process of the ontology among ontology engineering: The enterprise ontology and Uschold and King method (Uschold and Gruninger, 1996), The TOVE ontology and Gruninger and Fox method (Uschold and Gruninger, 1996), KACTUS and Bernaras method (Bernaras *et al*., 1996), CHEMICALS ontology and methontology method (Fernandez-Lopez *et al*., 1999) and The SENSUS ontology and method (Knight *et al*., 1995).

However, at present there is no standard and authority methodology in Ontology Engineering. Existing methodologies are born in specific projects and they are server for their projects. Due to different considerations of their respective areas and specific project, the process of build ontology is different. We can evaluate the maturity of methodology through the software life cycle method IEEE1074-1995 (Jing and Liansheng, 2004).

Thesaurus-based ontology construction: Many scholars presented to build ontology based on the existing thesaurus such as dictionaries, thesaurus and classification and they have tried in practice. Thesauri has clear semantic structures and can be used to extract concepts and relations. There are more than ten thesauri which have already been used to be converted into ontologies.

It can be divided into two categories according to the difference of construction method: One is to describe ontology directly based on thesaurus using the XML/RDFS syntax and does not adjust thesaurus. For example, Ven Eman (2005) put forward to describe thesaurus using OWL that is one of the description language of ontology. Jun (2003) put forward to define descriptors using RDFS to realize the conversion from thesaurus to ontology; The other one is ontology thought and it generate a new ontology through improving the thesaurus such as add or delete the concept of thesaurus, adjust the relations between concepts. For example, FAO convert Agrovoc into agricultural ontology using automatic ontology learning system through extracting relations between concepts (Kawtrakul *et al*., 2005). Qin and Paling (2001) convert the controlled vocabulary in GEM into ontology use Ontoligua system. Wiefinga *et al*. (2001) of University of Amsterdam uses the controlled vocabularies of Art and Architecture Thesaurus (AAT) to describe the ancient fruniture ontology.

Thesaurus and ontology are all build on the basis of comprehension of knowledge, so they have the premise of integration. But thesaurus and ontology are built for different objectives. As a standardized vocabulary terms, thesaurus can improve the retrieval efficiency of computer. Building the relationship between concepts is the core of ontology and the semantic content that the computer can understand is the premise of ontology. In order to convert thesaurus into ontology accurately, we need consider its characteristics, clean data and adjust semantic relation, so it will has some practical significance.

SEMANTIC QUERY EXPANSION

Most research related to ontology-based query expansion has focused on information retrieval in general search engines, but very few have been applied in the specific context of high-speed railway domain.

According to Chli and De Wilde (2006), query expansion can use methods such as lexical co-occurrence, clustering, stemming and knowledge models. Lexical co-occurrence is the process for establishing relationships between words based on the proximity analysis of the terms in a document. In clustering, documents that share a significant number of terms are grouped together in a cluster. The discriminant words from each cluster would be used to expand the query. In this case, it is assumed that similar documents are relevant to the same queries (Bhogal *et al*., 2007). Stemming is the process in which variations of terms are generated by the addition or removal of prefixes and suffixes as appropriate. This broadens or narrows the scope of the query. Finally, the knowledge-based methods extract the terms semantically related to the user's query (Navigli and Velardi, 2003). The knowledge-based methods in turn could be dependent or independent of the corpus (a collection of text balanced and annotated where the sense that the words take in a context is defined by the words that surround them).

Expanded terms may come from several sources which are dependent or independent of the corpus (Sartori, 2009). In the former case, the terms are obtained from document statistics. For example the frequency and the terms co-occurrence, or the similarity thesaurus (Croch and Yong, 1992). In other cases, the terms are obtained from external sources, e.g., models derived from expert knowledge or from the users' own recommendations (Navigli and Velardi, 2003).

Research by Huang and Hsu (2008) proposes a query expansion using a knowledge model based on the PubMed corpus complemented by Gene ontology's terms. The tree for the expansion adds the most relevant terms extracted from each abstract which also exist in the ontology. The terms' relevance is calculated by the formula tf idf where tf stand for term frequency and idf stand for inverse document frequency. The tree is created by levels; the first-level terms have the highest tf idf in the whole collection. For the second level, the tf idf is recalculated based on the collection associated with term that will be expanded. The terms selected for expansion are the most relevant terms, according to the idf and the level of depth. The experiment only assesses the validity of the extracted terms for expansion but does not evaluate the precision or recall of the results retrieved with the expanded queries.

Ontology-based query expansion is a method that extracts new terms from the knowledge represented in the concepts, attributes and relationships of an ontology. Methods for ontology query expansion differ in terms of the kind of ontology relationships used for expanding and the type of ontology used. For example, Ali and Khan (2008) propose a solution for query expansion based on a task ontology where the expansion is carried out using the following relationships: Synonym, specialization, lexical variant and acronym. The ontology-based framework proposes by Zou *et al*. (2008) consists of a domain ontology built for the proposal, the semantic annotation algorithm and semantic expansion reasoning algorithm. The expansion takes place using the part-of, kind-of or the instance-of relationships. In this case, it is important to note that the expansion algorithm evaluation is performed on tagged resources and indexed by the same ontology used in the expansion. While the proposal by Ali and Khan (2008) is interesting, we emphasise that the expansion algorithm evaluation is performed on an ontology created for the study, does not use a test collection and not listed the set of queries applied or the assessment of the results' relevance.

These researches have some similarities with our proposal, it is important to note that there are significant differences, including:

- The query object is the knowledge of high-speed railway that is comprised by a number of professional fields
- We considered a variety of extension mode includes synonymous-concepts expansion, sub-concepts expansion, upper-concepts expansion, relationship-concepts expansion and semantic similarity concepts expansion
- The results of semantic query sorted return to the user according to the similarity value between query expansion vector and document index vector

METHODOLOGY OF ONTOLOGY BUILDING FOR HIGH-SPEED RAILWAY DOMAIN

Some kinds of domains such as agriculture, railway, high-speed railway and aviation include different professional fields. For example, the high-speed railway domain consists of maintenance engineering, traction power supply, EMU and operation management. The existing method of ontology construction is not suitable for high-speed railway domain that is composed by different professional fields.

The ontology of high-speed railway domain can be integrated by ontologies of the professional fields that are built on thesaurus and thematic words. Because "thesaurus of railway" is not suitable for high-speed railway, so under the support of railway ministry of China we organize experts to rebuild "thesaurus and thematic words of high-speed railway" which is apply to high-speed railway literature information. Among these the first class category is ten that includes maintenance engineering, traction power supply, EMU, signal and control, communication, operation management, safety and rescue, environmental engineering, transportation economy and comprehensive evaluation, the second category under the first class category is 64, the third category under the second category is 208, 2488 subject headings corresponds these categories. At the same time we also build "foundation data table of high-speed railway" that includes 35 tables such as foundation data table for EMU, foundation data table for line and foundation data table for traction power supply etc. The first class category of "thesaurus and thematic words of high-speed railway" includes the professional fields of high-speed railway, so we can build ontologies of the professional fields based on "thesaurus and thematic words of high-speed railway". We can build properties and instances of the professional fields based on "foundation data table of high-speed railway".

We should clear that firstly the purpose of constructing high-speed railway domain ontology is to achieve semantic organization and semantic retrieval, secondly the hierarchical relation between concepts of domain ontology is selected based on "thesaurus and thematic words of high-speed railway" and use engineering thinking to construct domain ontology of high-speed railway. In the process of constructing ontology we should emphasize the participation of experts of high-speed railway domain. Because even if construct ontology through engineering method, it also needs expert to identify and evaluate.

According to the actual situation of high-speed railway domain and based on the two major methods ontology engineering and thesaurus-based ontology construction, this study puts forward a construction methodology of domain ontology that is composed of multi-disciplinary based on thesaurus and thematic words of high-speed railway that is built by the experts, shown as Fig. 1. First we build various majors ontology of high-speed railway and then they are preliminary merged into one unified domain ontology of high-speed railway. Building ontology of the professional field and merging them into one are important and the following details all parts of this two steps.

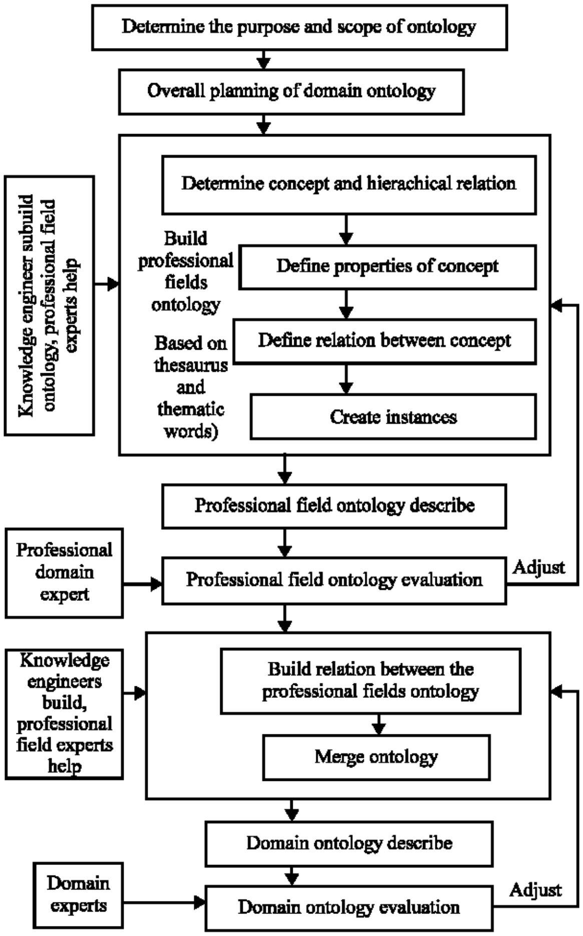
Determine the purpose and scope of ontology
Overall planning of domain ontology
Knowledge engineer subuild ontology, professional field experts help
Determine concept and hierachical relation
Build professional fields ontology
Define properties of concept
Based on thesaurus and thematic words)
Define relation between concept
Create instances
Professional field ontology describe
Professional domain expert
Professional field ontology evaluation
Adjust
Knowledge engineers build, professional field experts help
Build relation between the professional fields ontology
Merge ontology
Domain ontology describe
Domain experts
Domain ontology evaluation
Adjust

Fig. 1: Flow of domain ontology construction based on thesaurus and thematic words of high-speed railway

PRINCIPLE OF CONSTRUCTING ONTOLOGY FOR THE PROFESSIONAL FIELD

The model of ontology mainly includes concepts, properties, relations and instances (Aimin *et al*., 2005):

- **Choice concept:** The first class concept main comes from the words of "thesaurus and thematic words of high-speed railway". The second class concept is the classification of the first class concept and if the first class concept has two or more division methods, we often choice the most general division method as the second class concept. If the second class concept can still be classified, then divide it down and until no classification

For example, EMU is the first class concept of the ontology of EMU professional field. EMU has three division methods: According to the dynamic configuration, according to the fashion of supply power and according to speed grade. Among these the most commonly used is according to the fashion of supply power in which EMU is divided into "EMU of power distributed" and "EMU of central power". So, the second class concept of EMU is "EMU of power distributed" and "EMU of central power".

- **Determine the property:** The property of the professional field ontology mainly comes from "foundation data table of high-speed railway" and if the properties of the concept is not complete, that need the experts of high-speed railway domain to add

For example, we can add "type", "manufacture", "country" and "starting acceleration" properties to EMU concept according to "foundation data table of high-speed railway".

- **Determine the instance:** The instance in "foundation data table of high-speed railway" is not complete, so it needs domain experts to add. If the concept has sub-concept, the instance is added to the bottom concept

For example, we can add "CRH1 ", "CRH2", "CRH3" etc., instances to "EMU of power distributed" according to "foundation data table of high-speed railway".

- **Determine the relation:** There are two types of relations in ontology. One is hierarchical relationship including the hyponymy relation and instance relation and it can get from "thesaurus and thematic words of high-speed railway". The other is non hierarchical relationship and it can be described by adding the special property of concept to connect two concepts. Non hierarchical relationship need to be determined by domain experts. For example, add the special property of "supply power" to the concept of "pantograph" which can describe the non hierarchical relationship between the concepts of "pantograph" and "EMU". Therefore, we can get the relation: pantograph <supply power> EMU

MERGE THE PROFESSIONAL FIELD ONTOLOGIES

At present, we have constructed five professional fields's core ontology of EMU, operation management, safety and rescue, traction power supply and maintenance engineering. The thesaurus and thematic words of EMU includes four hundred and twenty one words. The depth of the EMU ontology network is 3. By now, we have defined forty five concepts, sixty and six properties, six user-defined relations and one hundred and ten instances. The thesaurus and thematic words of operation management includes one hundred and five words. The depth of the operation management ontology network is 3. By now, we have defined forty and eight concepts, thirty properties, one user-defined relation and twenty nine instances. The thesaurus and thematic words of safety and rescue includes thirty two words. The depth of the safety and rescue ontology network is 2. By now, we have defined nineteen concepts, seven properties, one user-defined relation and six instances. The thesaurus and thematic words of traction power supply includes three hundred and seventy eight words. The depth of the safety and rescue ontology network is 4. By now, we have defined 356 concepts, ten properties, five user-defined relations and 35 instances. The thesaurus and thematic words of maintenance engineering includes 989 words. The depth of the safety and rescue ontology network is 4. By now, we have defined 920 concepts, 25 properties, one user-defined relation and fourteen instances.

When the professional fields' core ontologies are built, we can merge the professional field ontology according to the relationship between different professional fields into one unified domain ontology of high-speed railway. For example, the concept of "contact network system" in traction power supply ontology associate with the concept of "pantograph" in EMU ontology through the relation of <supply power>.

DESIGN OF THE MODEL FOR SEMANTIC QUERY EXPANSION

Semantic similarity expansion is an important way to expand the user's retrieval words. Therefore, this section first study the semantic similarity algorithm between concepts of domain ontology, then discuss semantic query expansion model.

DOMAIN ONTOLOGY MODEL AND THE SEMANTIC SIMILARITY ALGORITHM BETWEEN CONCEPTS

Domain ontology model: At present, there are a lot of definitions about ontology. Among which Studer's explanation has received most agreement (Studer *et al*., 1998). In his theory, ontology is an explicit specification for the concept system and the definition includes conceptualization, explicit, formal and share. Ontology can be expressed by the modality as follows: O = {C, R, F, A, I}.

According to the definition and description of ontology, domain ontology reflects a common view of the given field. It describes the semantic information of concepts through defining the relation between concepts. In factual domain ontology, concepts have not only "is a" relation but also other relation. Especially in the domain ontology that is built with a lot of professional fields, concepts have much user-defined relation. So, it makes the structure of concept not just a complete tree-model but a network-model. Therefore, according to the character of domain ontology we rebuild the domain ontology model based on ontology model.

Definition 1: Domain ontology model is an Eight-tuples which can be stated as:

$$IDO = \{C, P, H^c, R^s, R^{ud}, I, F, A\}$$

Where IDO is domain ontology, C is concept or class, P is the Datatype property in domain ontology, H^c is the "subclass-of" relation of class, R^s is synonymy relation between class, R^{ud} is user-defined relation of classes (including "part-of" relation), that is, R^{ud} is the Object Property of class, I is the set of concept instance, F is an special relation between concepts which can be stated as follows: $c_1 \times c_2 \times .... \times c_{n-1} \rightarrow c_{n,}$, A is the axioms which domain ontology concepts or the relation between concepts meet.

Definition 2: Concept C is a Nine-tuples model which can be described as:

$$C = \{P, C_{sc}, C_{uc}, C_s, C_r, H_c, R_s, R_{ud}, I_c\}$$

Where P is Datatype property of concept C, C_{sc} is the upper class of concept C, C_s is equivalent class of concept C, C_r is the concept which is related to concept C, here mainly refers to the concepts that has user-defined relation with concept C, H_c is the "is-a" relation of concept C, R_s is the synonymy relation of concept C, R_{ud} is the user-defined relation of concept C, I_c is the instance of concept C.

There three types of concept relation:

- "Is-a" relation, denoted with $C_{sc,}$, C_{uc} and H_c
- Synonymy relation, denoted with C_s and R_s
- User-defined relation, denoted with R_{ud}

SEMANTIC SIMILARITY ALGORITHM BETWEEN CONCEPTS

The shortcoming of traditional ontology concept similarity computation is that it just takes hierarchy semantic relation into account, neglecting the influence of other semantic relation and the influence of object instance on concept (Zongze, 2006). This study proposes a new concept similarity computation model and it is built on definition 3. This model takes the influence of every element in ontology concept model on similarity into account. The elements include property (Datatype property), sub-class semantic relation, other semantic relation (user-defined) and character of instance, so the model is called Fourfold Matching-Distance Model (MD4).

- **"Is-a" relation semantic similarity computation:** This study considers the influence of the distance and levels of concept on concept similarity (Chen and Jiang, 2006). The equation is as follows:

$$\mathrm{Sim}_h(C_i, C_j) = \frac{\alpha \times (\mathrm{dl}(C_1) + \mathrm{dl}(C_2))}{(\mathrm{Dist}(C_1, C_2) + \alpha) \times 2 \times \mathrm{Maxdl} \times \max(|\mathrm{dl}(C_1) - \mathrm{dl}(C_2)|, 1)} \quad (1)$$

where, $dl(C_1)$ is the level of C_1 and $dl(C_2)$ is the level of C_2. $Dist(C_1, C_2)$ is the shortest path of concept C_1 and C_2 in ontology. Maxdl is the depth of ontology tree and it is divided in the equation is to ensure that the value of similarity is between 0 and 1. α is a adjustable parameter and is always bigger than zero.

- **Semantic similarity computation of concepts with user-defined relation:** The equation of user-defined relation similarity is showed as follows:

$$\mathrm{Sim}_{ud\_r}(C_i, C_j) = \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} \mathrm{Sim}(r_{udi}, r_{udj})}{\max(p, q)} \quad (2)$$

where, r_{udi} expresses p user-defined relations relevant to concepts C_i and r_{udj} expresses q user-defined relations relevant to concepts C_j.

We adopt Eq. 1 to compute the similarity of concepts which are relevant to user-defined relation. Then we get integrated similarity computation equation (Gao and Diao, 2011):

$$\mathrm{Sim}_{ud\_c}(C_i, C_j) = \frac{\sum_{i=1}^{m} \max(\sum_{j=1}^{n} \mathrm{Sim}_h(c_{ri}, c_{rj}))}{m} \quad (3)$$

where, C_{ri} expresses m concepts relevant to p user-defined relations and C_{rj} expresses n concepts relevant to q user-defined relations.

In domain ontology, the similarity of C_i and C_j reflected through user-defined relation is:

$$\mathrm{Sim}_{ud}(C_i, C_j) = \beta \mathrm{Sim}_{ud\_r}(C_i, C_j) + \gamma \mathrm{Sim}_{ud\_c}(C_i, C_j) \quad (4)$$

Where β and γ are weight (simple supposing $\beta = \gamma = 0.5$), $0<\beta<1$, $0<\gamma<1$, $\beta+\gamma = 1$.

- **Datatype property similarity computation:** The formula of property similarity computation for concept C_i and C_j:

$$\mathrm{Sim}_p(C_i, C_j) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Sim}(p_i, P_j)}{\max(m,n)} \quad (5)$$

The number of Datatype properties of C_i is m and C_j's is n.

- **Instance semantic similarity computation:** Instance semantic similarity computation is the same with the computation of Datatype property. The instance semantic similarity computation equation for C_i and C_j is as follows:

$$\mathrm{Sim}_i(C_i, C_j) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{Sim}(i_{ci}, i_{cj})}{\max(m,n)} \quad (6)$$

- **Factual similarity computation of concepts which are not synonymous in domain ontology:** Through integrating the four kinds of similarity, we get the factual semantic similarity computation equation of C_i and C_j which are not synonymous in domain ontology:

$$\mathrm{Sim}_{ud}(C_i, C_j) = \omega \mathrm{Sim}_h(C_i, C_j) + \theta \mathrm{Sim}_p(C_i, C_j) + \mathrm{Sim}_{ud}(C_i, C_j) + \beta \mathrm{Sim}_i(C_i, C_j) \quad (7)$$

where, ω and θ are weight, $0<\omega<1$, $0<\theta<1$, $\omega+\theta = 1$, normally, ω is bigger.

BASIC DEFINITION OF THE SEMANTIC QUERY

Extended retrieval is through searching the related concepts in search terms to find the related knowledge. The followwing are the definitions of semantic relatedness and query expansion.

Definition 3: Domain ontology defines the vocabulary that composed of the field and relationships between them. In ontology, we say the C_1 and C_2 is semantic relatedness if there is a path Path (c_1, c_2) between node C_1 and C_2.

We make $Q(\{C_i\})$ indicates the query of concept or concept set C_i, i = 1, 2, 3,.... $Set(Q(\{C\}))$ indicates the result set of query and it is denoted by Set(Q) or $Set(C_i)$.

Definition 4: We say the query to C_2 is the extension of the query to C_1 if $Set(C_1) \subseteq Set(C_2)$ and we say the query to C_2 is the semantic extension of the query to C_1 if C_1 and C_2 is semantic relatedness.

As can be seen from the above definitions, the key of extension retrieval is to find the related concepts.

Definition 5

Set of sub-concepts: The sub-concepts set of concept t is $L(t) = \{l \in T | l$ is the sub-concept of concept $t\}$. T is the set of concepts. In the following T is the same meaning.

Definition 6

Set of upper-concepts: The upper-concepts set of concept t is $U(t)=\{u \in T \mid u$ is the parent concept of concept $t\}$.

Definition 7

Set of synonymous-concepts: The synonymous-concepts set of concept t is $E(t)=\{e \in T \mid e$ is the synonymous concept of concept $t\}$.

Definition 8

Set of relationship-concepts: The relationship-concepts set of concept t is $R(t) =\{r \in T | r$ is the concept that is connected with concept t through relationship$\}$.

MODEL OF THE SEMANTIC EXTENSION

This study only considers that the user's retrieval words are related to the ontology concepts.

According to the definition in section 4.2, this study gives five kinds of semantic expansion methods: Synonymous-concepts expansion, sub-concepts expansion, upper-concepts expansion, relationship-concepts expansion and semantic similarity concepts expansion.

Domain ontology model is stored by the Owl file.

In order to facilitate the realization of semantic retrieval, we define search function as follows:

FS(c) = $\{synonymy | synonymy \in DO\}$: obtain the synonymous concepts
FC(c) = $\{child | child \in DO\}$: obtain the sub concepts

FP(c) = {parent|parent∈DO}: obtain the upper concepts
FR(c) = {relation|relation∈DO}: obtain the concepts that is connected with the concept through relationship
FSS(c) = {semantic similarity concepts(SSC)|SSC∈DO}: obtain the concepts according to the similarity between the user's retrieval words and ontology concepts

Model of synonymous-concepts expansion: This model is to expand the synonymous concepts of the retrieval words.

Algorithm 1: Model of synonymous expansion

```
Input:   The user's retrieval words
Output:  The result set of synonymous concepts or NULL
Begin
Step 1:  Judeg if the retrieval words entered by the user are concepts, if not,
         exit the algorithm
Step 2:  Initialize the result set $S_{set}$, set it is null and load domain
         ontology model (.owl) file
Step 3:  Executive FS(c) function, if the return is not empty and then put
         the result into $S_{set}$
Step 4:  Return the result set $S_{set}$
End
```

The algorithm of sub-concepts expansion, upper-concepts expansion and relationship-concepts expansion are same with algorithm 1, so we omit them in the study.

Model of semantic similarity concepts expansion: This model is to expand the concepts according to the similarity between the user's retrieval words and ontology concepts. Because the size of domain ontology is bigger, so in order to improve the efficiency of the algorithm, we compute the semantic similarity between concepts in prior according to the Eq. 7 and the similarity are stored in ontology knowledge base.

Algorithm 2: Model of semantic similarity expansion

```
Input:   The user's retrieval words
Output:  The result set of semantic similarity concepts or NULL
```

ARCHITECTURE OF SEMANTIC RETRIEVAL SYSTEM BASED ON SEMANTIC QUERY EXPANSION

We adopt JSP to develop our system and adopt Protégé that is developed by Stanford University to build our domain ontologies. Shown as Fig. 2, our proposed system is consists of four main modules: Ontologies building, retrieval words semantic extension, semantic precomputation for document and semantic retrieval (Li *et al*., 2011). Domain ontology is stored in Ontologies Library and document knowledge that has been indexed by domain ontology is stored in database. The following are some brief descriptions about these four parts.

ONTOLOGIES BUILDING

Building ontology demands standardized expression fashion and work steps. Describing language and modelling mode are chosen according to the level of ontology, the use of ontology, the principle of building ontology and evaluation criteria. In our system we use OWL describing language for ontology. OWL is a standard of ontology describing language in semantic web that is recommended by W3C and it is an extension on the basis of RDFS (Kong *et al*., 2008).

Method about building ontology is not mature and there is not has a complete and unified methodology. However, a number of typical ontology and methodology have been developed. In our system we adopt our method proposed in section 3 to build various professional fields ontology and they are preliminary merged into one unified domain ontology of high-speed railway.

The depth of high-speed railway domain ontology network is 4. Under which we compute the semantic similarity between concepts in prior according to the Eq. 7 and the similarity are stored in ontology knowledge base. At the same time, the number in document base of literature that is related with high-speed railway is more than thirty thousands. Figure 3 is a segment of high-speed railway domain ontology.

RETRIEVAL WORDS SEMANTIC EXTENSION

The main function of this module is that: Semantic expand the retrieval words terms entered by the user. We use the method in section 4 to expand the retrieval words entered by the user and use one-dimensional vector form to represent the expanded concepts and their corresponding weights. The following is the concept vector and weight vector:

$$\text{Document}(1) = \{c_{21}, c_{22}, \ldots, c_{2n}\} \quad (8)$$

$$\text{Weight}(1) = \{w_{21}, w_{22}, \ldots, w_{2n}\} \quad (9)$$

SEMANTIC PRECOMPUTATION FOR DOCUMENT

This module achieves documents formalization. We index document in prior by ontology concepts in ontology knowledge base. Benefiting from the ontologies that we have built, document formalization can use the

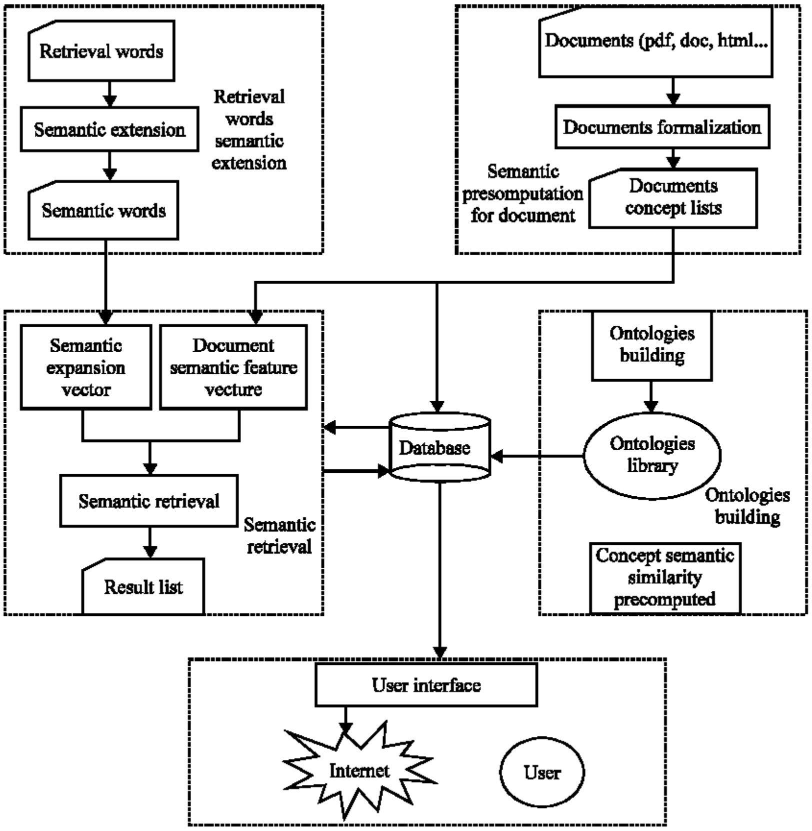

Fig. 2: Architecture of semantic retrieval system based on semantic query expansion

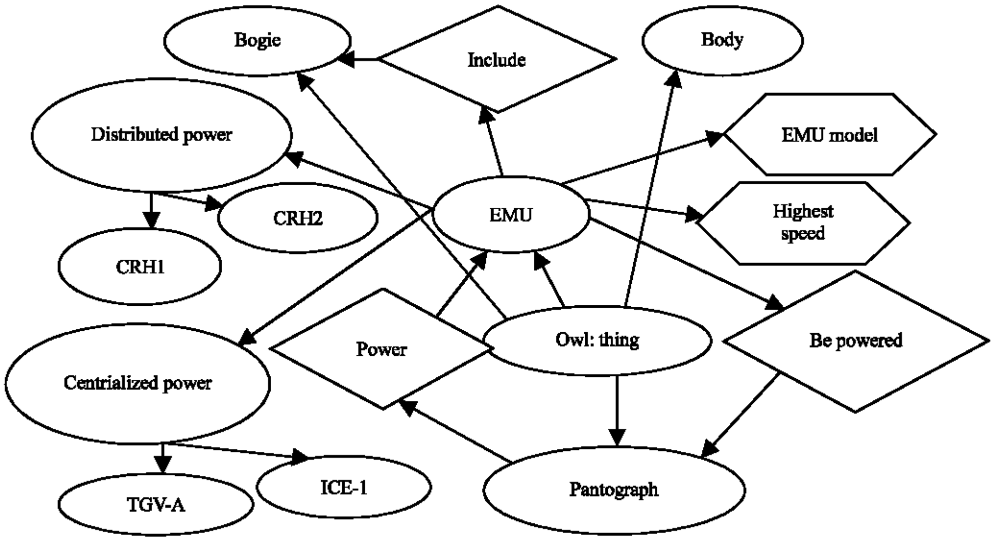

Fig. 3: Segment of high-speed railway domain ontology

concept to formalize the documents containing information. The input data of this module can be various kinds of files, including PDF, Doc, HTML, XML and so on while the output data is a list of concepts derived from the ontologies (Sanchez *et al*. 2011).

The goal of indexing document is to obtain the document's semantic vector (De Macedo *et al.*, 2009). How to compute the weight of indexing concept is the key. Previous studies showed that the frequency and location plays an important role in reflecting the relationship between indexing words and the theme of document. In this study we adopt the method combined nonlinear function and "paired comparison" to compute the weight of indexing concept that is proposed by Zheng and Lu (2005). We use one-dimensional vector form to represent the concept of indexing documents and their corresponding weights. The following is the concept vector and weight vector:

$$\text{Document}(2) = \{c_{11}, c_{12}, \ldots, c_{1m}\} \tag{10}$$

$$\text{Weight}(2) = \{w_{11}, w_{12}, \ldots, w_{1m}\} \tag{11}$$

SEMANTIC RETRIEVAL

The main function of this module is to calculate the similarity between semantic extension vector that is expressed by Eq. 8 and 9 and indexing vector of document that is expressed by Eq. 10 and 11. Then compare the similarity to the threshold that is established by the user and if it is bigger than the threshold, then the document is relevant to the user's query. Finally, the document list that is sorted according to the similarity is returned to the user interface.

We use the method of Wu and Yang (2005) to compute the semantic similarity between the document vector and the semantic extension vector. First calculate the semantic similarity between two concepts and then calculate the semantic similarity between two vectors.

The corresponding weight of concept c_{1i} in Document(1) is w_{1i} and the corresponding weight of concept c_{2j} in Document(2) is w_{2j}. The formula to compute the similarity between c_{1i} and c_{2j} is:

$$\text{sim}(c_{1i}, c_{2j}) = -\frac{\log\left(\frac{w_{1i} + w_{2j}}{2}\right)}{\text{distance}(c, c_{1i}) + \text{distance}(c, c_{2j}) + 1} \tag{12}$$

where, distance (c_1, c_2) is the number of edges of the shortest path between c_1 and c_2.

The formula of calculating the similarity between the two one-dimensional vector is:

$$\text{sim}(\text{Document}(1), \text{Document}(2)) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \text{sim}(c_{1i}, c_{2j})}{m \times n} \tag{13}$$

where, m is the number of concepts in Eq. 8 and n is the number of concepts in Eq. 10.

EXPERIMENTATION AND EVALUATION

We make an experiment on high-speed railway domain ontology and analyze the effectiveness of semantic extended retrieval.

The evaluation of an information retrieval system often concerns Recall and Precision based on a reference collection which consists of a set of testing documents, a set of example queries and a set of relevant documents chosen by specialists for corresponding queries. Let R be the set of relevant documents for a given information retrieval request and A be the answer set of documents searched by the retrieval system. The recall and precision are defined as follows respectively:

$$\text{Recall} = \frac{|A \cap R|}{|R|} \times 100\% \tag{14}$$

$$\text{Precision} = \frac{|A \cap R|}{|A|} \times 100\% \tag{15}$$

We make ten times experiments to compare the results with semantic retrieval method (Expand) in this paper and traditional keyword search method (Non_expand). Figure 4 shows the different Recall and Precision.

We can see that the result of semantic retrieval method (Expand) is prior to the result of traditional keyword search method (Non_expand). So, it proved that the semantic extended retrieval method is effective atsome extend. Although the relevant set of artificial selection has some uncertainty but this uncertainty can not be completely avoided in the human-computer interaction systems.

Recall	Precision		
Non_expand (%)	Expand (%)	Non_expand (%)	Expand (%)
40.46	49.10	30.65	40.72
41.01	44.75	32.33	38.94
33.56	52.92	40.00	50.40
40.86	48.96	41.60	58.68
43.69	59.75	45.90	65.70
48.30	53.50	49.56	62.45
47.80	54.90	50.78	69.26
50.10	51.10	42.34	52.90
42.60	50.60	39.52	54.50
51.50	60.30	47.56	59.50

Fig. 4: Experiment results

CONCLUSION

Semantic web based on ontology provides a new guidance way for information retrieval. Domain ontology can be used to extend the user's retrieval words and index domain knowledge. This study discusses the semantic query expansion using domain ontology constructed by thesaurus and thematic words for High-speed Railway Knowledge. Firstly, according to the actual situation of high-speed railway domain that is comprised of different professional fields, this study puts forward a construction methodology of domain ontology based on thesaurus and thematic words of high-speed railway that is built by the experts. Guided by this method, we build various majors ontology of high-speed railway and then they are preliminary merged into one unified domain ontology of high-speed railway. Secondly, this study designs the model of semantic query expansion based on domain ontology and gave five kinds of semantic expansion methods. Then the model of semantic query expansion is applied in semantic retrieval system. Thirdly, we design a semantic retrieval system and the model of semantic query expansion is applied in it. The designed system consists of four main components: Ontologies building, Retrieval words semantic extension, semantic precomputation for document and semantic retrieval. Finally, this study makes a comparison between traditional keyword search method and our semantic retrieval method. The experimental results show that the developed semantic retrieval system can improve the recall and precision.

In our future work, we will perfect the domain ontology and consider the evolution problem of domain ontology of high-speed railway. In the semantic retrieval system of high-speed railway knowledge, ontology reasoning retrieval is our next work.

ACKNOWLEDGMENTS

This study was financially supported by the Natural Science Foundation of Hebei Province of China (F2012208018, F2013208107), the Project of Science and Technology Department of Hebei Province of China (134576113D), the Project of Hebei University of Science and Technology (QD201017), National Natural Science Foundation of China (51271033, 71301044) under which the present study was possible.

REFERENCES

Aimin, T., Z. Zhen and F. Jing, 2005. Thesaurus-based approach to build domain ontology. New Technol. Library Inform. Serv., 122: 1-5.

Ali, W. and S. Khan, 2008. Ontology driven query expansion in data integration. Proceedings of the 4th International Conference on Semantics, Knowledge and Grid, December 3-5, 2008, Beijing, China, pp: 57-63.

Bernaras, A., I. Laresgoiti and J. Corera, 1996. Building and reusing ontologies for electrical network applications. Proceedings of the 12th European Conference on Artificial Intelligence, August 11-16, 1996, Budapest, Hungary, pp: 298-302.

Bhogal, J., A. Macfarlane and P. Smith, 2007. A review of ontology based query expansion. Inform. Process. Manage., 43: 866-886.

Chen, J. and Z.H. Jiang, 2006. Concept similarity computation for domain ontology. Comput. Eng. Appl., 33: 163-166.

Chli, M. and P. De Wilde, 2006. Internet search: Subdivision-based interactive query expansion and the soft semantic web. Applied Soft Comput., 6: 372-383.

Croch, C.J. and B. Yong, 1992. Experiments in automatic statistical thesaurus construction. Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, June 21-24, 1992, Copenhagen, Denmark, pp: 77-87.

De Macedo, D.D.J., A. Von Wangenheim, M.A.R. Dantas, H.G.W. Perantunes, 2009. An architecture for DICOM medical images storage and retrieval adopting distributed file systems. Int. J. High Perform. Syst. Archit., 2: 99-106.

Fernandez-Lopez, M., A. Gomez-Perez, J. Pazos-Sierra and A.P. Sierra, 1999. Building a chemical ontology using methontology and the ontology design environment. IEEE Intell. Syst. Appl., 14: 37-46.

Gao, H. and M. Diao, 2011. Cultural firework algorithm and its application for digital filters design. Int. J. Modell. Infertility. Control, 14: 324-331.

Huang, Y.F. and C.H. Hsu, 2008. PubMed smarter: Query expansion with implicit words based on gene ontology. Knowledge-Based Syst., 21: 927-933.

Jin, Z., 2001. Ontology Research in Knowledge Engineering. In: Knowledge Engineering and Knowledge Science, Chang, S.K. (Ed.). Tsinghua University Press, Beijing, pp: 447-465.

Jing, L. and M. Liansheng, 2004. Comparison of seven approaches in constructing ontology. New Technol. Library Inform. Serv., 112: 17-22.

Jun, M., 2003. Research on RDF-based thesaurus. J. China Soc. Sci. Tech. Inform., 22: 163-168.

Kara, S., O. Alan, O. Sabuncu, S. Akpýnar, N.K. Cicekli and F.N. Alpaslan, 2012. An ontology-based retrieval system using semantic indexing. Inform. Syst., 37: 294-305.

Kawtrakul, A., A. Imsombut, A. Thunyakijjanukit, D. Soergel and A. Liang *et al*., 2005. Automatic term relationship cleaning and refinement for agrovoc. Proceedings of the 5th Conference of the European Federation for Information Technology in Agriculture, Food and Environment, July 25-28, 2005, Vila Real, Portugal..

Klein, M. and A. Bernstein, 2001. Searching services on the semantic: Web using process ontologies. Proceedings of the International Semantic Web Working Symposium, July 30-August 1, 2001, California, USA., pp:159-172.

Knight, K., I. Chander, M. Haines, V. Hatzivassiloglou and E. Hovy *et al*., 1995. Filling knowledge gaps in a broad-coverage machine translation system. Proceedings of the14th International Joint Conference on Artificial Intelligence, August 20-25,1995, Montreal, Canada, pp: 1390-1396.

Kong, T.Y., W.L. Li and H.O. Zhang, 2008. Research in medicine ontology-based information retrieval system. J. Hebei Univ. Sci. Technol., 29: 223-226.

Li, Y., F. Guo and X. Wang, 2011. Intelligent personalised information retrieval system based on multi-agent. Int. J. Modell. Infertility. Control, 12: 113-118.

Navigli, R. and P. Velardi, 2003. An analysis of ontology-based query expansion strategies. Proceedings of the Workshop on Adaptive Text Extraction and Mining, September, 2003, Dubrovnik-Croatia. pp: 42-49.

Qin, J. and S. Paling, 2001. Converting a controlled vocabulary into an ontology: The case of GEM. Inform. Res., Vol. 6.

Qiu, Y. and H.P. Frei, 1993. Concept based query expansion. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, June 27-July 01, 1993, New York, NY., USA., pp:160-169.

Sanchez, D., M. Batet and D. Isern, 2011. Ontology-based information content computation. Knowledge-Based Syst., 24: 297-303.

Sartori, F., 2009. A comparison of methods and techniques for ontological query expansion. Proceedings of the 3rd International Conference on Metadata and Semantic Research, October 1-2, 2009, Milan, Italy, pp: 215-225.

Segura, N.A., Salvador-Sanchez, E. Garcia-Barriocanal and M. Prieto, 2011. An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the Gene ontology. Knowledge Based Syst., 24: 119-133.

Studer, R., V.R. Benjamins and D. Fensel, 1998. Knowledge engineering: Principles and methods. Data Knowl. Eng., 25: 161-197.

Uschold, M. and M. Gruninger, 1996. Ontologies: Principles, methods and applications. Knowl. Eng. Rev., 11: 93-136.

Ven Eman, J., 2005. Owl exports from a full thesaurus. Bull. Am. Soc. Inform. Sci. Technol., 32: 22-26.

Wiefinga, B.J., A.T. Schreiber, J. Wielemaker and J.A.C. Sandberg, 2001. From thesaurus to ontology. Proceedings of the 1st International Conference on Knowledge Capture, October 21-23, 2001, Victoria, BC., Canada, pp: 194-201.

Wu, J. and G. Yang, 2005. An ontology-based method for project and domain expert matching. Proceedings of the 2nd International Conference Fuzzy Systems and Knowledge Discovery, August 27-29, 2005, Changsha, China, pp: 176-185.

Zheng, J. and J. Lu, 2005. Study of an improved keywords distillation method. Comput. Eng., 31: 194-196.

Zongze, W., 2006. A new fast rate control algorithm for JPEG2000. Int. J. Modell. Infertility. Control, 1: 159-163.

Zou, G., B. Zhang, Y. Gan and J. Zhang, 2008. An ontology-based methodology for semantic expansion search. Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery, October 18-20, 2008, Jinan Shandong, pp: 453-457.