Research on Semantic Retrieval for Communication Ontology

Fang Wenting, Li Yunqing, Xiong Yanlong , Chen Jing, Yan Xiaopeng

Jiangxi Normal University, Nanchang, Jiangxi, 330022, China
wshczmj@163.com

***Abstract*—Seven-step method is used to build the communication domain ontology, and the system uses Jena stored the ontology to the relational database. It Uses ICTCLAS Chinese word segmentation machine to deal with retrieval request submitted by users and converted to specification format which the system can identify. Using distance-based semantic similarity algorithm which adds edge type for quantizing semantic distance between concepts, this algorithm improve the accuracy of quantitative the semantic similarity between concepts, and building a semantic retrieval for communication ontology system.**

Keywords-Ontology; Communications; Semantic Retrieval

I. INTRODUCTION

With the development of society, the information data grow rapidly, which means that the advent of the big data era. How to get more accurate and effective use information in the digital age has become the issue which people more and more attention.

The current information retrieval can be divided into three categories: full text retrieval, data retrieval and semantic retrieval[1]. Full text retrieval and data retrieval are traditional. Full text retrieval is based on user-submitted search terms just to find matching each word in the full text, without considering the semantic information degree of the matching degree between the user entries submitted and the concepts in the text. Data retrieval requires that the query mode and the data stored have consistent format and fixed structure. These two kinds of retrieval methods have some limitations, do not have the ductility and adaptability. Although full text search and data retrieval are simple and fast, but can't find inner link between information, the retrieval results can not accurately reflect the real needs of the user. Semantic retrieval refers to a new retrieval model that develops based on the original retrieval model, it is retrieving accurate the result that the user wishes by the semantic relationships between the keywords and other keywords, not limited to the literal meaning of the user’s submits, can be more accurately meet the needs of the most and return the results to the user.

Ontology has a good concept hierarchy , it can be well representation of semantic relationships and semantic relationships between concepts, such as synonymous relationship, hyponymy relationship and part-of relationships. Ontology is the core technology of semantic retrieval system, ontology can solve the semantic descriptions and ambiguous of the concept problem.

II. ONTOLOGY THEORY AND CONSTRUCTION

A. Ontology Theory

Ontology is a philosophical concept at earliest, ontology refers to the interpretation or explanation of the objective existence system, it concerns with the abstract nature of objective reality. Ontology generally used the definition is Gruber in 1993 "An ontology is an explicit specification of a conceptualization."[2]. Studer gives a ontology definition based on Gruber’s definition "ontology is a shared conceptual model an explicit formal specification". The definition of ontology includes four layer of meaning: conceptualization, explicit, formal and share.

Constitute of ontology can be given by using the following formula[3]:

Ontology =Concept + Property + Axiom+ Value + Norminal

Among them, the concept can be divided into two kinds: original concept and definition concept. For example, "The tree is a plant" is the original concept, "There is a 90 degree angle triangle is a right triangle" is the definition concept. Property is a description of the concept characteristics or nature. Axiom is defined on the conceptual property constraints and rules. Value is a specific assignment. Nominal is the concept has no example, or the instance is used in the concept definition.

B. Ontology Construction

Ontology construction is a huge project, which requires domain experts or professional and technical person in accordance with thesauri and other expertise to build ontology through the ontology construction tool. Existing ontology construction method is based on a particular aspect of the application, which requires the participation of domain experts, it belongs semi-automated ontology construction. At present, ontology can not be automated build. Ontology construction method is mainly divided into two categories: ontology engineering methods and the thesaurus convert into a ontology methods. This paper adopts to the method of converting the thesauri to ontology.

The main foreign construction methods have IDEF5 Method, ENTERPRISE Method, TOVE Method, METHONLOGY Method, KACTUS Method, Seven-step Method and SENSUS Method[4]. According to the maturity level arrangement in turn is the Seven-step Method, METHONLOGY Method, IDEF5 Method, TOVE Method, ENTERPRISE Method, SENSUS Method and KACTUS Method.

978-1-4673-7644-0/15 $31.00 © 2015 IEEE
DOI 10.1109/ICICTA.2015.192

III. SEMANTIC RETRIEVAL

A. Semantic Retrieval

The traditional retrieval is based on keyword retrieval, the input keyword and the keyword in the repository are according to the character match directly, this retrieval method does not take into account the semantic relationship between keywords. Semantic Retrieval gets keyword match by using semantic relations between keywords in the ontology, so the retrieval results not only include the results of the traditional retrieval methods, but also include the intrinsic semantic relevant results. The results of semantic retrieval are more comprehensive and accurate.

Research and build model of semantic retrieval system is the main research areas of ontology-based semantic retrieval. The current semantic retrieval research focuses on the semantic processing of information resources in order to achieve higher retrieval efficiency, extraction and processing of semantic information can be based on semantic web methods and techniques, can also be based on the technology of natural language processing, the former is relatively more common in the semantic retrieval research[5].

B. Semantic Similarity

Semantic similarity is refers to the degree of similarity between the two concepts, usually refers to a common feature of the two concepts have. The main use of hyponymy connected (is-a) constitute the concept of hierarchy when calculating the concept similarity. There are other relations in addition to the is-a relationship between the two concepts, such as the whole-part (part of) relationship. When constructing communication ontology in addition to these two relationship and the equal relationship, it means the two keywords is a same concept..

Semantic similarity algorithm is divided into three categories: distance-based semantic similarity algorithm, content-based semantic similarity algorithm and attribute-based semantic similarity algorithm[6]. The idea of distance-based semantic similarity algorithm is quantified the semantic distance through the path length of two concepts in ontology tree classification system, when the semantic distance between two concepts is greater, the similar degree is lower, when the semantic distance is smaller, the similar degree is higher. Wu and Plamer algorithm and Leacock and Chodorow algorithm are the representative algorithms. The basic principle of content-based semantic similarity algorithm is: the shared information of two concepts is proportional to the semantic similarity between them. Lord et al.[7] put forward to calculate the semantic similarity between different concepts for using informational content contained in the public parent node, they calculate the similarity between the compared concepts for direct using the information of the most recent parent node. The content-based semantic similarity algorithm puts the sharing information content between the concepts and their common parent node as the information content of the concepts. The idea of attribute-based semantic similarity algorithm is the more public property between the two concepts, the semantic similarity is greater, because the concept express by its characteristic properties. Tversky algorithm[8] is a typical algorithm of the attribute-based semantic similarity algorithm, the attribute-based semantic similarity algorithm neither considers the semantic distance between the concept words, nor considers the information content of the concept words and their common parent node, but only considers their attribute attribute information.

The distance-based semantic similarity algorithm assumptions each side has the same weight when calculates the semantic distance, without taking into account different relationship of the side has different weight. There are several types of ontology side, such as equal relationship and hyponymy relationship, the side of the different types has different weight. If sub concept and parent concept are equal relationship then the edge weight is 1, if sub concept and parent concept are hyponymy relationship then the weight is 1/2, if sub concept and parent concept are other relationship then the weight is 1/3, specifically as follows.

$$t=\begin{cases}1 & \textit{equal relationship}\\ 1/2 & \textit{hyponymy relationship}\\ 1/3 & \textit{other relationship}\end{cases} \tag{3-1}$$

Semantic similarity of two concepts c1 and c2 is calculated as follows:

$$Sim(c1,c2)=\frac{2\sum_{i\in(1,n)} t_i}{\sum_{j\in(1,m)} t_j+\sum_{k\in(1,p)} t_k+2\sum_{i\in(1,n)} t_i} \tag{3-2}$$

$\sum_{i\in(1,n)} t_i$ is the shortest distance from the nearest common parent node of c1 and c2 to the root node, $\sum_{j\in(1,m)} t_j$ is the shortest distance from c1 to the nearest common parent node, $\sum_{k\in(1,p)} t_k$ is the shortest distance from c2 to the nearest common parent node.

IV. COMMUNICATE DOMAIN ONTOLOGY SEMANTIC RETRIEVAL SYSTEM

A. Construction Communication Domain Ontology

This paper uses seven-step method, a higher maturity, to build ontology. The following steps are used to build ontology :

(1) Determine the area of expertise and knowledge ontology category. This paper builds communication domain ontology, and determines the scope of the field of the communication through the "Chinese Thesaurus".

(2) Investigate the possibility of reuse of existing knowledge ontology. We don't consider reuse other ontology, directly to build ontology.

(3) List the important terms of ontology. Abstracting terms in the field of communication accords to "China's Thesaurus", and obtaining the relationship between keywords from the thesaurus, the relationship as follows: Y (formal thesaurus), D (informal thesaurus), F (lower thesaur-

us), S (upper thesaurus), Z (family of the first word) and C (related thesaurus).

(4) Define the class and the class level system. In this paper, define the class hierarchical relationship for using the keyword layer thesauruses. Courses in the field of communications are divided into professional core course, radio and television direction, the direction of mobile communication and other professional courses. Professional core courses include information theory and coding theory, communication theory, television theory, the electromag-netic field and the electromagnetic wave propagation, anten-na and radio. Radio and television transmit direction includ-es digital TV technology, radio and television broadcasting and digital sending technology. The direction of mobile communication includes mobile communications, modern switching technology and mobile TV technology. Other professional courses include circuit analysis, linear electronic circuits, nonlinear electronic circuits, digital circuits, signals and systems, digital signal processing, SCM principles and interface technology and computer networks. The subclass also contains a lot of concepts, such as communication theory subclass also contains news, information, the amount of information, channel, noise source, analog signals, digital signals, bandwidth, base band, linear modulation, nonlinear modulation, source coding, channel coding, frequency distortion, modulation and demodulation and so on.

(5) Define the class property. Ontology model consists of object property and data property. The object property describes the relationship between the different property of the class or the instance. There are equal relations (equality), upper keyword (Upper subject headings), lower keyword (Under a subject), in which the upper and lower is contrary notion that A is B's upper, then B is A's lower. A keyword can has multiple upper keywords, may be also have multiple lower keywords. For example, telecommunications mathematics' upper keywords are communication and applied mathematics, communication networks' equal relations are telecommunications network and network, communication networks' lower keywords are fax net, telephone network, the public communications network, computer communication network, communication subnets, synchronous communication network, satellite network, user communications network, power communication network, multimedia communication networks, broadband communication networks, digital communications networks and micro-channel communications network. The data property has title, keywords, press, author, and other Chinese books DOI.

(6) Custom facets of the property. The facets of the property is refers to the value of the type, scope, number and other characteristics about attribute values. For example, the value type of Chinese Library Classification number is the character and the scope is the Chinese Library Classification. The instance (i.e. the book) can only take a value, but the subject heading can take a value or multiple values.

(7) Create the instance. The instance in this ontology is the book in the field of communication, and adds the object property and data property values. Such as communication principle includes the books like communication principle (Aijing Sun,Wei Dang and Liping Ji compile), communication principle(Changxin Fan and Lina Cao compile) and simple tutorial of communication principle. So the class of communication principle includes the instances like communication principle (Aijing Sun,Wei Dang and Liping Ji compile), communication principle(Changxin Fan and Lina Cao compile) and simple tutorial of communication principle.

We use protégé 4.3 to build the ontology which is developed by Stanford University, protégé can be simple, visual and directly on the class hierarchy build the ontology, protégé can be directly storage into the OWL, does not need to input code.

B. Data Storage

As the library has a huge amount of books, uses the OWL file storage of ontology is not conducive to directly inference, query, and retrieve from the ontology, therefore we should convert the ontology into a database to store. Then we create a mapping between the field of communication database and in the field of ontology knowledge database, When a query request coming, it can be achieved by mapping between the two databases, and associated information retrieval based on ontology. In the process, the mapping between the ontology and the database using the following rules to achieve:

Rule 1: The class in ontology is corresponding to the base table in database;

Rule 2: The instance in ontology is corresponding to the table records in database;

Rule 3: The OWL data types is corresponding to the table data types in database, as shown in Table 1.

Table 1. OWL and database data types correspond

Data types	Data types in OWL	Data types in database
The basic numeric	xsd: decimal	DECIMAL
	xsd: integer	INT
	xsd: long	LONG
	xsd: short	SMALLINT
Character	xsd: string	CHAR/TEXT
Time to date	xsd: data	DATE
	xsd: time	TIME

C. Participle

Keywords entered by the user may be not a single word, so it's necessary to divide the input word. ICTCLAS full name is Institute of Computing Technology, Chinese Lexical Analysis System. It's a segmentation system studied by

Chinese Academy of Sciences. The main idea of this system is word segmentation through by CHMM (cascading shape Markov model), through the word segmentation, both increases the accuracy of the segmentation, and ensures the efficiency of word segmentation. ICTCLAS is based on C / C ++ development, has java interface, supports C / C ++ / C # / Delphi / Java and others mainstream development languages. ICTCLAS' word segmentation results can be directly labeled the speech, you can remove the "and" "mean," "the," and other banned word segmentation based on the results, the system can be converted into a recognizable pattern specification. Table 2 is the results of the ICTCLAS segmentation word.

Table 2. the results of the ICTCLAS segmentation word

The words before statement	The results after participle
Communication domain ontology semantic retrieval	Communications / v domain / n ontology / n of / u semantic / n retrieval / v

Segmentation results on the table uses the dictionary from ICTCLAS system. The description of ontology is used the form of a triple for storage.

D. Semantic Retrieval

The function of semantic retrieval module is making semantic expansion for user input query, through studies of the semantic extension(equal relationship, hyponymy relationship and the other relationship) of keywords to get results, the result is mapping to the ontology concepts terminology which extension from semantic relations between keywords, and finding relevant examples information of these terms in the instance library, and bring up the appropriate library information from the matching result. Specific implementation structure of Semantic Retrieval module as follows.

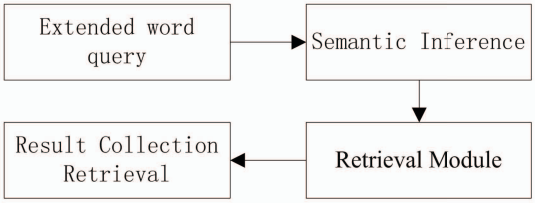

Figure 1. The structure of Semantic Retrieval module

(1) Extended word query is for keywords collection which get from participle and processing the user input query to semantic expansion.

(2) Semantic inference is a key part of the semantic retrieval module. Jess is a low-level and general-purpose inference engine. Jena is a semantic Web oriented application development package, it includes a more comprehensive content, In addition to containing the inference engine, it also supports ontology's resolve. Jena is the ontology inference engine, Jena inference engine is not an expert inference engine, it adds a field shell on the basis of Jess inference engine not a inference engine design expert. The accuracy and efficiency of Jena is better than generic inference engine just for the Web ontology language inference. This article uses Jena inference engine for semantic inference.

The rule of inference engine provided by Jena supports the forward inference, the backward inference and the hybrid inference. Jena has many internal reasoning rules, these rules are mainly defined for ontology, they are mainly used for transferring the relation and the property of different classes. Here are two internal rules.

[rdfs2: (?x ?p ?y), (?p rdfs:domain ?c)-> (?x rdf:type ?c)]

This rule instructions that if the range value of p is c, c is a instance of the class x.

[rdfs9: (?x rdfs:subClassOf ?y), (?a rdf:type ?x) -> (?a rdf:type ?y)]

This rule instructions that if the class x is a subclass of the class y and a is a instance of the class x, a is a instance of the class y.

Jena supports user-defined rules, the rule as follows.

[rule1:(?x fa:Under a subject?y)(?x fa:zushouis ? z)->(?y fa:zushouis ?z)]

[rule2:(?x fa:Upper a subject?y)(?x fa:zushouis ?z)->(?y fa:zushouis ?z)]

Rule1 instructions that if the under word of the concept x is the concept y and the family head of the concept x is the concept z, the family head of the concept y is the concept z.

Rule2 instructions that if the hypernym of the concept x is the concept y and the family head of the concept y is the concept z, the family head of the concept x is the concept z.

Inference the user's query keywords is related to the concept collection of uses Jena and the concept in the concept collection to sort on of uses semantic similarity algorithm. Because of the distance-based semantic similarity algorithm does not consider different edge has different importance, content-based semantic similarity algorithm puts the sharing Information content between the concepts and their common parent node as the information content of the concepts, attribute-based semantic similarity algorithm only considers the property does not considers other factors, so we use distance-based semantic similarity algorithm which adds edge type for quantizing semantic distance between concepts in the concept collection and sorts the concept collection according to the quantitative results.

(3) Retrieval module finds the appropriate book information in the relational database based on the keywords collection of semantic reasoning. Semantic retrieval uses SPARQL query language, SPARQL queries and retrieves ontology by matching triples graphical pattern because of ontology stores information in accordance with triples (Subject, Predicate, Object). SPARQL achieves the match query by replacing query variable to the corresponding RDF vocabulary.

(4) Search result collection is taking the retrieval module retrieves the results in order and according to the sorted concept collection.

E. Experimental results

The system provides two retrieval methods: traditional keyword-based retrieval and ontology-based semantic retri-

eval, the experiment retrieves results by two retrieval methods.

Entering the "communication network". Inputting "communication network" in the traditional keyword-based retrieval system, this system retrieves book information only includes the "communication network" word, like communication network, public communication network, multimedia communication network and digital communication network. Traditional retrieval results show in Figure 3.

Figure 2. Traditional retrieval results

Inputting "communication network" in the semantic retrieval system, semantic inference module can infer that the telecommunication network and communication network are identical relationship, these two words represent the same concept , and fax network, telephone network, public communication network, computer communication network, communication subnet, synchronous communication network, satellite network, user communications network, power communication network, multimedia communication network, broadband communication network, digital communication network and micro-channel communication network are the lower keywords of communications network. Querying and retrieving the results based on Inference results. Outputting the book information about all the synonyms and lower keywords. Semantic retrieval results show in Figure 3.

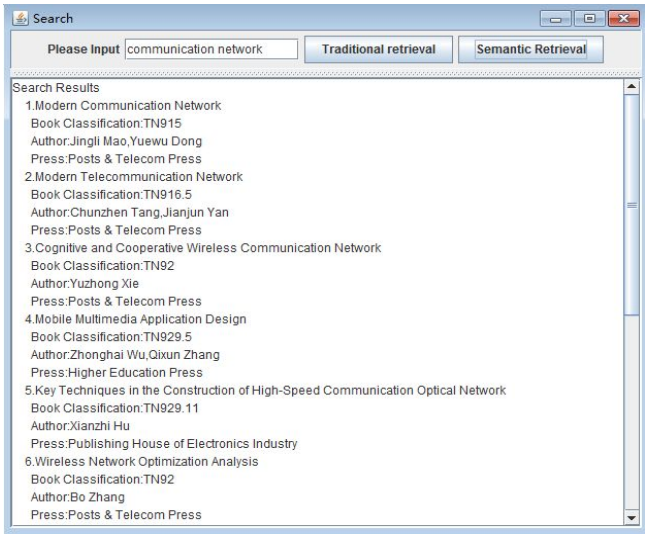

Figure 3. Traditional retrieval results

Although the retrieval result of traditional retrieval won't miss any information which matches the key character, but the traditional retrieval couldn't retrieve the result which semantic related with the keyword. Semantic retrieval neither misses the result which matches the key character nor misses the result which semantic related with the keyword. Experiments show that the recall and precision of semantic retrieval system are better than traditional keyword retrieval.

V. CONCLUSIONS

In this system, we build the communication domain ontology and use ICTCLAS word segment user-entered keywords, based on the segmentation results, we do some semantic extension, reasoning and retrieval. This system shows that Semantic Retrieval is better than traditional keyword in recall and precision rates. As we build communication domain ontology library in this paper is based on thesaurus constructed without experts, so it may be not quite perfect.

REFERENCES

[1] Xiyu Wang, Zhonglin Zhou. Construction Digital Library Information Retrieval Model Based on Ontology[J]. Information Research, 2011(9):21-23.

[2] Gruber T R.A Translation Approach to Portable Ontology Specifications[J]. Knowledge Acquisition, 1993(5):199-200.

[3] Wei Shong,Ming Zhang. Concise Guide to the Semantic Web[M]. Higher Education Press, 2004.

[4] Binbin Yu. Research on Ontology Construction Methods and Build Tools[J]. The Border Economy and Culture, 2012(12): 167-168.

[5] Liyun Kang, Xiaoyue Wang, Rujiang Bai. Econometric analysis of domestic research on semantic retrieval[J]. Journal of Modern Information, 2012(5):104-109.

[6] Hong Zhe, De Xu. Summary of Ontology-based Semantic Similarity and Relevance Calculation[J]. Computer Science, 2012(2):8-13.

[7] Lord P W, Stevens R D, Brass A, et al. Investigating Semantic Similarity Measures across the Gene Ontology: The Relationship Between Sequence and Annotation [J]. Bioinformatics, 2003,19(10): 1275-1283.

[8] Tversky A. Features of Similarity [J].Psychological Review, 1977,84(4): 327-352.