Search Engine System Based on Ontology of Technological Resources

Weihui Dai
School of Management, Fudan University, Shanghai 200433, P.R.China
Email: whdai@fudan.edu.cn

Yu You, Wenjuan Wang, Yiming Sun and Tong Li
School of Software, Fudan University, Shanghai 200433, P.R.China,
School of Software,Yunnan University, Kunming 650091
School of Software, Fudan University, Shanghai 200433, P.R.China,
School of Software, Yunnan University, Kunming 650091, P.R.China
Email: 063053214@fudan.edu.cn, wenjuan42@gmail.com, 09212010017@fudan.edu.cn, tli@ynu.edu.cn

***Abstract*—Internet has become a huge and updating information warehouse, and provides a new source for us to build a well technological resources sharing system to support our research work and development activities. However, the technological resources on Internet is usually diverse, professional and complex. They are difficult to be retrieved precisely and completely by traditional search engines. This paper proposed a new search engine system based on ontology of technological resources. In that system, a database with ontology knowledge warehouse was designed to store all related conceptions and the relationships of technological domains. By semantic analysis of users' queries and a heuristic search, the expected technological resources can be retrieved more precisely and completely to satisfy their intentions.**

***Index Terms*—search engine, technological resources, ontology, semantic analysis, query intention, knowledge warehouse**

I. INTRODUCTION

With the rapid development of network and computer technology, Internet has become a huge and changing information warehouse, and Web has become a main way in which people get information. People tend to get information via different search engines, such as Google, Baidu, Sina, and Yahoo. But it is difficult to get the precise information from the multitudinous network data, because most of the traditional search engines adopt the search technology based on keywords full-text matching, which has shortness in understanding users' query intention. In the information global era, research and development activities are very important to improve our competition ability, while technological resources sharing is good support for research and development activities. So it is essential for us to have the intelligent search engine, which can understand natural languages and analysis users' query intention just like human being by semantic analysis.

This research was supported by National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).
Corresponding author: Weihui Dai

In data sharing service network, different domain's data may call each other and lots of technological resources come from different channels, such as news, newspapers, websites, or different countries, and these are different in description languages, forms and other aspects. When users input keywords to query information, after search engine searches the web-page indexing database, there may be lots of relevant web-page links, but little is really needed information. Such is so-called "Rich data, Poor information".

Technological resources are a set of human resources engaged in technological activities, material, financial, management, information and other hardware and software. It includes not only the instruments, equipment, but also experimental materials, experimental methods, experimental data and scientific talents. Network information, data, literature and other technological resources have been broadly shared, and it is an effective way to improve the efficiency of scientific research that try to make full use of technological resources and sharing.

Although there are abundant of technological resources on Internet, they are usually: large amount of data, complex types, different in store structures, isolated from each other.

In order to solve the technical bottleneck and satisfy users' search demands, this paper proposed a framework of search engine based on ontology of technological resources, semantic analysis, and data sharing, which is different from the traditional search engine in following aspects: (1) It can analyze the semantic of users' query conditions to improve the efficiency and precision. (2) The database includes all related conceptions and relationships of technological domain as query conditions. (3) It can filter the key information as feedback search results.

The rest of this paper is organized as follows: Section 2 introduces the search engine and ontology structure. Section 3 describes the design of search engine based on ontology. Section 4 discusses the implementation of such search engine. And section 5 is the conclusion.

doi:10.4304/jsw.6.9.1729-1736

Ⅱ. SEARCH ENGINE AND ONTOLOGY STRUCTURE

In recent years, the traditional search engine cannot satisfy users' higher query demands, because of its shortness. There are three limitations for traditional search engine based on keywords matching: (1) So much relevant information, and users cannot find out the precise information they need quickly. (2) Users cannot express their intentions only by inputting some keywords. (3) Poor semantic analysis, because the search engine cannot understand natural language. For example, when a user input "apple" as one of keywords, the search results are including both the information of a kind of fruit and a brand of computer.

In order to solve the limitations of traditional search engine, lots of researches try to optimize the algorithm and apply mode of search engine. However, the proposal and application of ontology and semantic web has become a hot issue of the new generation search engine.

A. Search Engine

Search engine is a kind of search tool designed to help users search for information in network. Search engines work in following steps: (1) Web crawling: search engines work by storing information about many web pages, which they retrieve from the html itself. These pages are retrieved by a web crawler (sometimes also known as a spider) — an automated Web browser which follows every link on the site. (2) Indexing: when users enter query keywords into a search engine, the engine examines its indexing and provides a listing of best-matching web pages according to its criteria. The indexing is built from the information stored with the data and the method by which the information is indexed. (3) Searching: the usefulness of a search engine depends on the relevance of the result set it returns back.

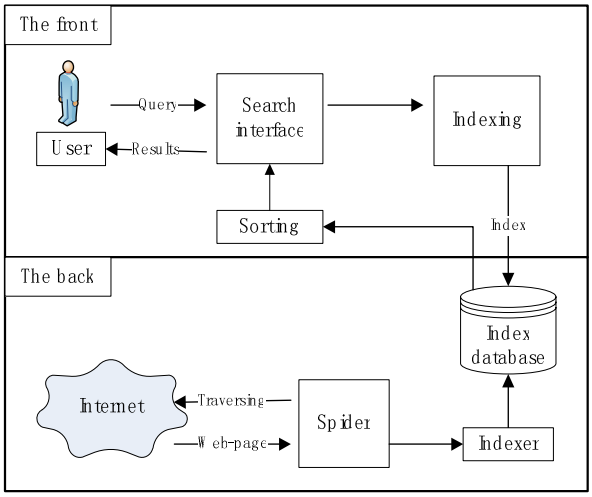


Figure 1. The structure of search engine

In fact, what the search engine searches is not the whole network, but the web-page database which stores lots of available data. The search results are usually presented in a list of results, which may consist of web pages, images, information and other types of files [1]. Figure 1 shows the structure of a search engine. There are three modules in the framework of search engine system: collecting module, indexing module, and retrieval module [2].

B. Ontology Structure

Ontology consists of vocabulary and a set of constraints on the way terms can be combined to model a domain [3] [4]. And now the most popular definition of Ontology is proposed by Gruber [5], namely "Ontology is the explicit standard explanation about conceptual model"[6]. There are four characters of ontology: explicit, formalization, sharing, and conceptualization. And the ontology consists of five elements: class, relationship, function, axiom, and instance.

Ontology knowledge warehouse is the core database of search engine building, and ontology is the basic absolute cell of ontology knowledge warehouse. The structure of ontology is as following:

O=<C, P, R, I>.

O represents Ontology.

C represents Classes, which is also named Concepts, and it is the collection of objects.

P represents Properties, which is the description of concepts.

R represents Relations of classes. There are four kinds of relations between classes: part of, kind of, instance of, and attribute of.

I represents Instances, which is the instantiation of concepts.

There are many kinds of ontology description language, such as RDF, RDFS (RDFSchema), OIL, DAML, OWL, KIF, SHOE, XOL, OCML, Ontolingua, CycL, and Loom. The ontology description language should satisfy following rules [7] [8]: (1) a well-defined syntax, (2) a well-defined semantics, (3) efficient reasoning support, (4) sufficient expressive power, (5) convenience of expression.

The popular ontology building methodologies are: (1) Skeletal Methodology [9], (2) TOVE [10], (3) METHONTOLOGY [11], (4) IDEF-5 methodology, (5) Seven-step methodology. The key rules of ontology design are sharing and reuse. Ontolingua and Protégé are popular ontology editors.

Ⅲ. DESIGN OF SEARCH ENGINE SYSTEM BASED ON ONTOLOGY

A. Functions and Workflow

The most difference between the traditional search engine and the search engine based on ontology of technological resources is that the former just matches keywords and database indexing, while the latter conducts the semantic analysis according to keywords firstly, and then carries through search processing.

The key point of our research is building an ontology module of technological resources and a semantic analysis module for this system. In the whole system, the core of data layer is ontology knowledge warehouse, and the core of business layer is semantic analysis. To improve the rate of recall and precision of the search engine system, a well-designed semantic analysis module

is important. Ontology technique is most popular in semantic analysis research, while ontology service is also the basic and key part in language system. We can describe the definitions and relations of objects, and realize the relevancy of conceptions, which can help computers reason and analyst users' query intention. Thus, computers can understand the nature language as human beings.

The design of search engine focuses on technological resources, with respect to the character of this domain, which is: (1) High value, (2) Resources sharing, (3) High risk of management, (4) Abundance of resources.

In order to satisfy users' search demands of technological resources, integrating nature language understanding, ontology technique, and relevant data mining technique, our system tries to implement an easy operating interface on Web platform, and feeds back the best results according to keywords users input in. Figure 2 shows the work-flow of search engine system.

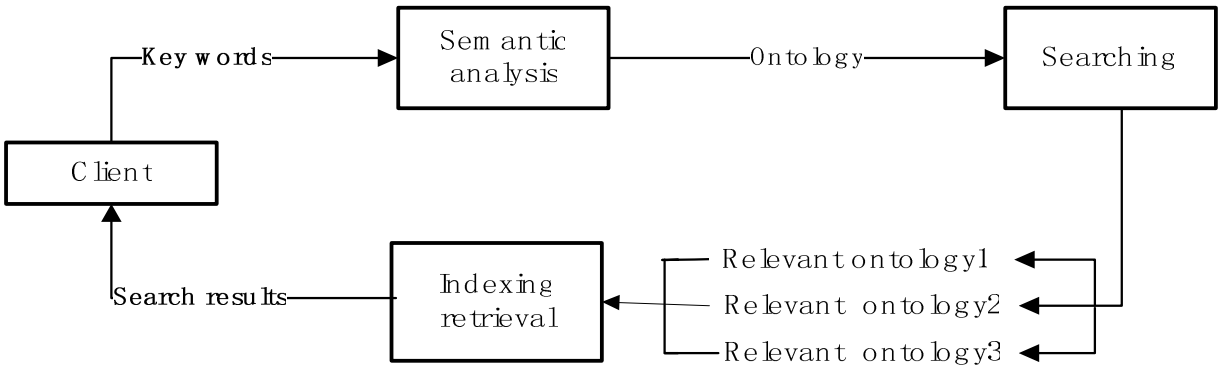


Figure 2. Work-flow of search engine system

The searching steps of search engine are as following:

(1) Users input some keywords in a kind of nature language on search page.

(2) The system conducts semantic analysis of keywords firstly, and then forms an ontology of the issue.

(3) The system searches other related items of the specific ontology according to the relationship of ontologies.

(4) The system retrieves technological resources database with the relevant ontology of step (3) as new keywords, therefore it can obtain the related technological resources.

(5) Finally, the system feeds back searching results by sorting, organizing, and a series of other processing.

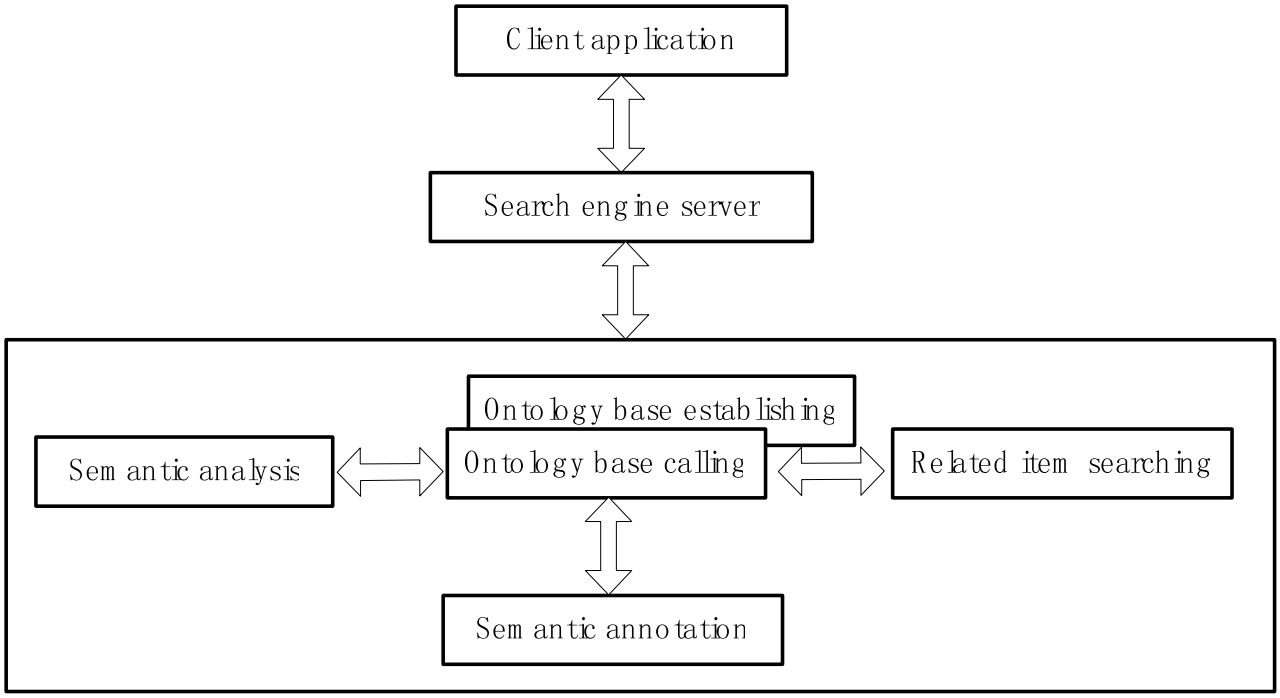


Figure 3. The architecture of search engine system

B. Architecture and Structure

As shown in the work-flow of search engine system, the system interprets keywords into the description of standard word library, which can be understood by computer according to the uniform conception description of ontology. Figure 3 represents the architecture of search engine system.

The search engine calls the pre-established technological resources ontology base firstly, which is the basis of data processing during semantic analysis, related item searching, and indexing.

Figure 4 shows the layer structure of search engine. It can be classified four layers by function: Presentation layer, Business layer, Data layer, and Object layer.

(1) Presentation layer: Presentation layer is user interface. It accepts users' input, and shows the output.

(2) Business layer: Business layer is also named service layer. It interconnects all the core processing modules, and enables them to exchange service information and requests.

(3) Data layer: the ontology base includes ontology conceptions, sub-classes, instances, and relations. The system will establish technological resources indexing according to the ontology base, and then establish technological resources indexing database.

(4) Object layer: the search engine is designed based on the scene of Shanghai research and development public service platform system. And technological resources objects include scientific data, documentation paper, instrument, specimen, and so on. All the information of these technological resources needs to be digitalized firstly, and then the key information will be stored in technological resources database.

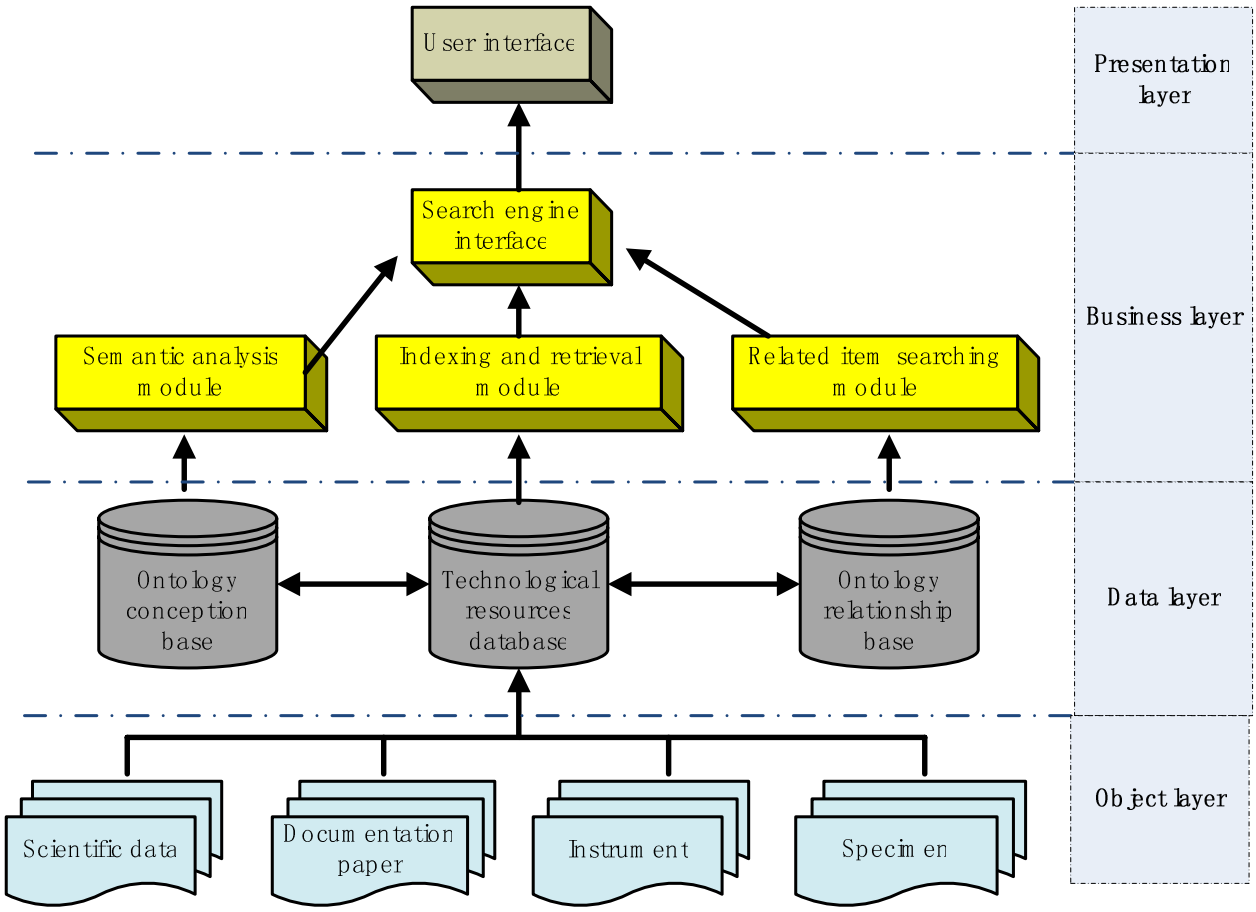


Figure 4. The layer structure of search engine system

C. Operation Model

The operation model of search engine system is as Figure 5, and Figure 6 is the implementation framework based on the design model of Figure 5.

Functions of four core modules are as following:

(1) Query preprocessor: it converts users' requests according to the ontology, and it can convert nature language query words into computer-readable information by semantic analysis.

(2) Demand converser: it can convert query words into the keywords that can be used to retrieve directly.

(3) Indexer: it indexes the keywords, and matches with the pre-established science and technological information base, then outputs the matching information.

(4) Annotator: it annotates kinds of technological resources, and then the right technological resources will be stored in database.

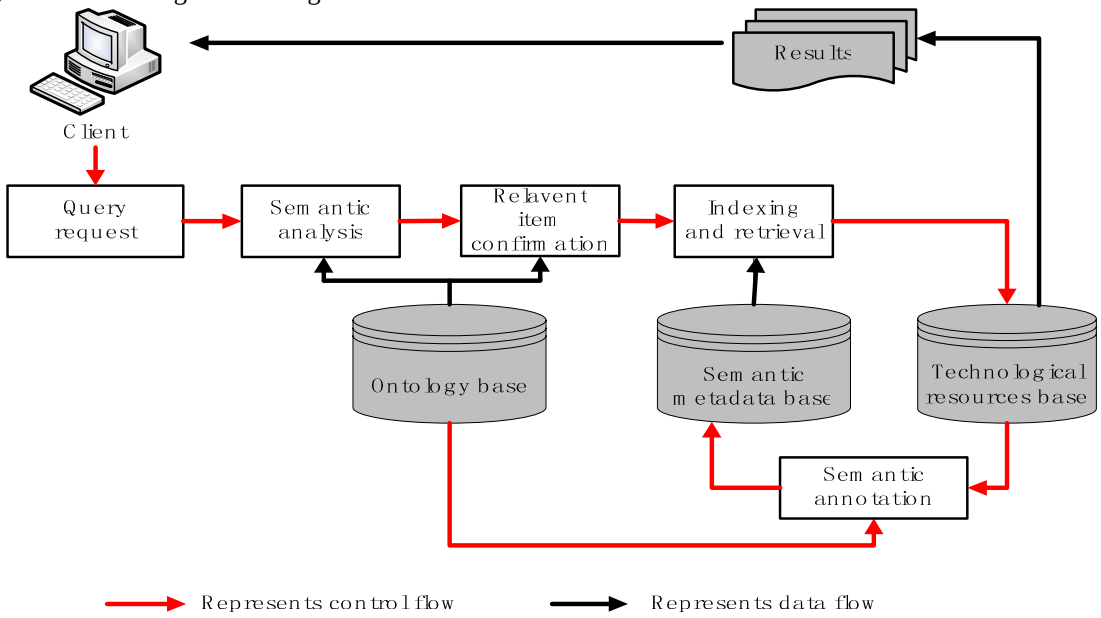


Figure 5. The operation model of search engine system

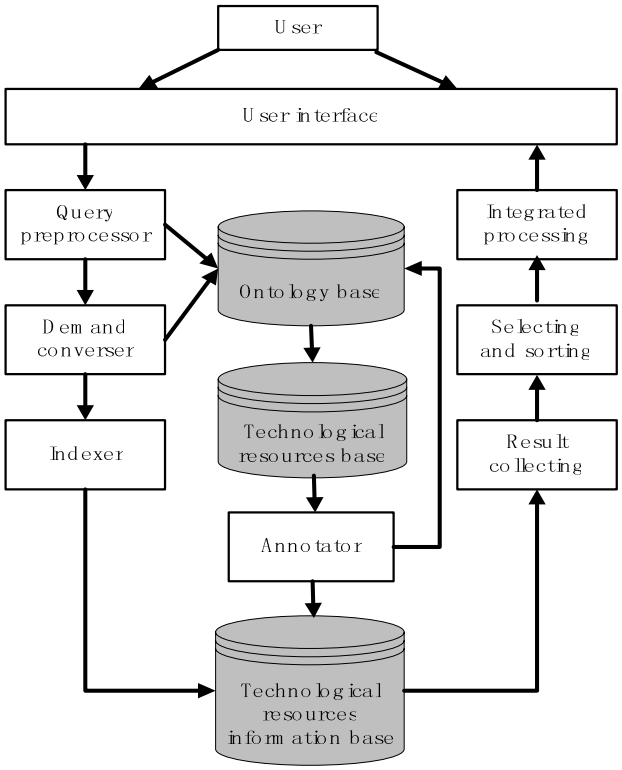


Figure 6. The implementation framework

In Figure 6, Ontology base is the core module of the system, which is called by Query preprocessor, Demand converser, and Indexer. Technological resources base stores the technological resources of scientific data sharing service of Shanghai research and development public service platform system. Technological resources information base is a digitalized technological resources base, which stores the annotate resources.

Users input their query demands by user interface, and then the search engine analysts the semantic and confirms the relevant items according to ontology base. After indexing and retrieval processes based on semantic metadata base, the search engine feeds back the most relevant information as search results to users.

Ⅳ. IMPLEMENTATION OF SEARCH ENGINE SYSTEM BASED ON ONTOLOGY

In the technological search engine system, Client interface is developed based on Web technique, and it adopts XHTML, JavaScript, and JSP. Jena API is used here to implement semantic analysis and relevant item confirmation processing, because the information is stored as OWL (Web Ontology Language) type ontology files. Technological resources are stored in technological resources base, OWL type ontology files are stored in ontology base, and semantic metadata base stores the description information of technological resources base.

A. Description of Technological Resources

In technological resources database, some objects may have the same name or attribute, so it is a good choice that “one object, one ID”, and this procedure is finished by annotator. Take the MASEP-SRRS Gamma Knife for example, and its ID is Exact_Instrument_17 in the Ontology. The definition of MASEP-SRRS Gamma Knife as follows:

```
<Word rdf:ID="Exact _Instrument_17">
  <Instrument:hasName>
    <Instrument:Name rdf:datatype="http://www.w3.org/2001/XMLSchema#string"> MASEP-
SRRS Gamma Knife </Instrument:Name>
  </Instrument:PlaceTo>
  <Instrument:PlaceTo>
    <Instrument:Address rdf:datatype="http://www.w3.org/2001/XMLSchema#string">No. 12,
Rd. Wulumuqi, Shanghai</Instrument:Address>
  </Instrument:PlaceTo>
  <Instrument:OwnBy>
    <Instrument:Owner rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Gamma
branch, Fudan attached Huashan hospital </Instrument:Owner>
  </Instrument:OwnBy>
  <Instrument:MadeFor>
    <Cancer rdf:ID="Cancer_35">
     <Instrument:Cure rdf:resouce="Exact _Instrument_17"/>
     <Cancer:Name rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Stomach
cancer</Cancer:Name>
    </ Cancer >
  </Instrument:MadeFor>
</Word>
```

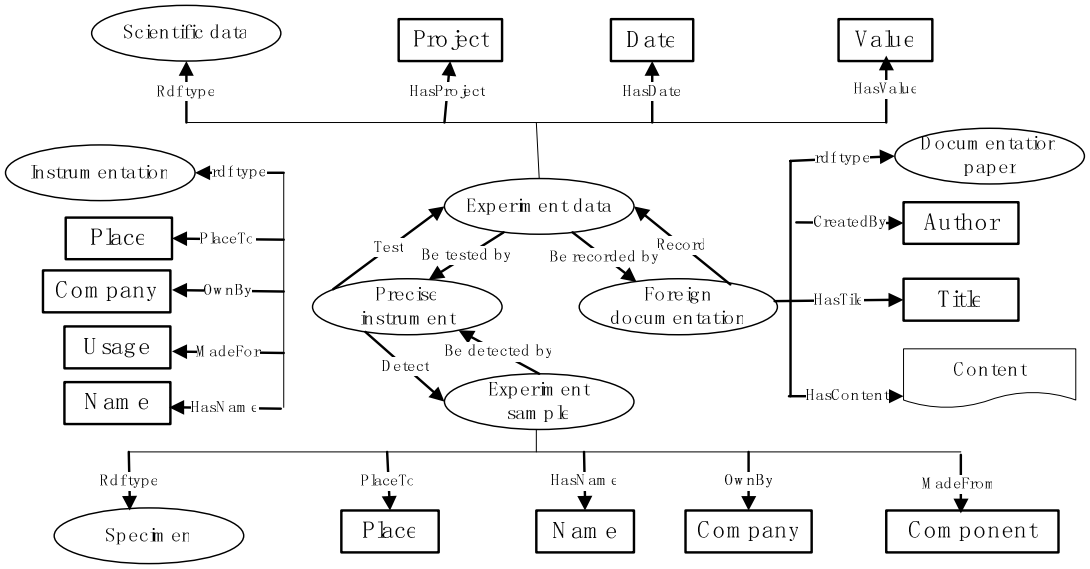


Figure 7. The non-formalized ontology model segment

B. Establishment of Ontology

We adopt Skeletal Methodology to establish the ontology of technological resources, and there are five steps:

(1) Confirm the subject domain of ontology.
(2) Create ontology model.
(3) Ontology formalization.
(4) Ontology evaluation.
(5) Ontology maintaining.

Figure 7 is the non-formalized ontology model segment.

C. Semantic Analysis of Query Request

Semantic analysis module is responsible for converting users' keywords into the conceptions of ontology, and inferring new indexing keywords from the selected domain.

It is common that different users input different keywords to describe the same object, or maybe in different languages. So it is important to use ontology-based approach to understand the keywords in nature language. Here, we take Chinese search engine as a case study. For example, "胃癌" and "胃上皮组织恶性肿瘤" in Chinese refer to the same thing with "stomach cancer" in English.

According to the conceptions inheritance relationship of ontology base, the description of ontology conception "stomach cancer" like this:

```
<Select case Keywords
  Case “胃癌” Return C = “ Stomach cancer”
  Break;
  Case “胃上皮组织恶性肿瘤” Return C= “Stomach cancer”
  Break;
  Case “Stomach cancer” Return C=“Stomach cancer”
  Break;
</Select>
```

According to users’ query keywords, indexing domain, and the relationship between ontologies, the search engine can query ontology model, which is a RDF (Resource Description Framework) triple [12]. Here, we adopt RDQL (RDF Data Query Language) as query language. RDQL is a query language for RDF in Jena models. RDF provides a graph with directed edges — the nodes are resources or literals. RDQL provides a way of specifying a graph pattern that is matched against the graph to yield a set of matches. It returns a list of bindings — each binding is a set of name-value pairs for the values of the variables. RDQL includes five kinds of sub-clauses: SELECT clause, FROM clause, WHERE clause, AND clause, and USING clause.

The following representation uses RDQL to query the relationship between ontology, with “stomach cancer” as the keywords and “instrument”

```
<Select ?instrument,?Exact_instrument_ID,?cancer_ID
 WHERE (?cancer_ID:hasName"Stomach cancer ")
       (?cancer_ID:BeCureBy,?Exact_instrument_ID)
       (?Exact_instrument_ID:hasName,?instrument)
 USING cancer FOR http://127.0.0.1/cancer#
       instrument FOR http://127.0.0.1/instrument#
       rdf FOR http://www.w3.org/2000/05/09-rdf-syntax-ns#
</Select>
```

According to RDQL query language, we can get the visual query graph as Figure 8. And the query result is shown in table 1.

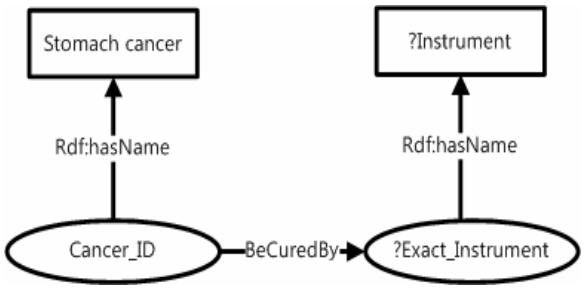


Figure 8. Ontology relationship query graph

By relationship converting of keywords “stomach cancer” in “instrument” domain, user’s query demand is changed into “Gamma Knife” searching demand, realizing semantic reasoning, finding the relevant technological resources at the same time.

TABLE 1.
QUERY RESULT

Cancer_ID	Cancer_35
Exact_Instrument_ID	Exact_Instrument_17
Instrument	Gamma Knife

D. Experiment and Results

The basis of technological resources base indexing is ontology classification. Our search engine is based on the technological resources of Shanghai research and development public service platform, and the search engine focuses on Chinese searching. Figure 9 shows the directory of Gamma Knife ontology. Figure 10 represents the search results as feedback.

Figure 9. Directory of Gamma Knife ontology

Figure 10. Search results

V. CONCLUSION

This paper researched the search engine system based on ontology of technological resources, and it also described the design and implementation in details. Compared with traditional search engine, it can improve the precision of search results by semantic analysis, satisfying users' search demands at the same time.

However, the search engine also has some shortness, such as it takes longer time to finish searching, and it is hard to maintain because of lacking of standard. With the development of semantic Web, this search engine system can be extended to multi-application and multi-platform, and it also can be implemented in other languages.

ACKNOLEDGMENT

This research was supported by National High-tech R & D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).

REFERENCES

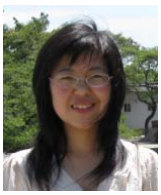
[1] http://en.wikipedia.org/wiki/Web_search_engine.
[2] Y. Khopkar, A. Spink, C.L. Giles, P. Shah, and S. Debnath, "Search engine personalization: an exploratory study," *First Monday*, vol.8(7), pp.1-23, 2003.
[3] http://www.isi.edu/isd/KRSharing/vision/AIMag-small.html.
[4] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, and W. Swartout, "Enabling technology for knowledge sharing," *AI Magazine*, vol.12(3), pp.36-56, 1991.
[5] T.R. Gruber, "A translation approach to portable ontologies," *Knowledge Acquisition*, vol. 5(2), pp.199-220, 1993.
[6] Z.Q. Du, J. Hu, H.X. Yi, and J. Z. Hu, "The research of the semantic search engine based on the ontology," *Proceedings of Wireless Communications, Networking and Mobile Computing*, pp. 5403-5406, 2007.
[7] Borst, W. N., *Construction of Engineering Ontologies for Knowledge Sharing and Reuse*, Enschede: University of Twente, 1997.
[8] http://www.w3.org.
[9] M. Uschold and M. Gruninger, "Ontologies principles, methods and applications," *Knowledge Engineering Review*, vol.11(2), pp.1-63, 1996.
[10] M. Gruninger and M.S. Fox, "Methodology for the design and evaluation of ontologies," *Workshop on Basic Ontological Issues in Knowledge Sharing*, IJCAI-95,1995.
[11] M. Ferndndez, A. Gomez-Perez, and N. Juristo, "Methontology: From ontological art towards ontological engineering," *Spring Symposium on Ontological Engineering*, 1997.
[12] http://www.w3.org/RDF/.

Weihui Dai received his B.S. degree in Automation Engineering in 1987, his Master degree in Automobile Electronics in 1992, and his Ph.D. in Biomedical Engineering in1996, all from Zhejiang University, China.

He worked as a post-doctor at School of Management, Fudan University from 1997 to 1999, a visiting scholar at Sloan School of Management, M.I.T from 2000 to 2001, and a visiting professor at Chonnam National University, Korea from 2001 to 2002. He is currently an Associate Professor at the Department of Information Management and Information Systems, School of Management, Fudan University, China.

Dr. Dai has published more than 120 papers in Software Engineering, Information Management and Information Systems, Complex Adaptive System and Socioeconomic Ecology, Digital Arts and Creative Industry, etc. Dr. Dai became a member of IEEE in 2003, a senior member of China Computer Society in 2004, and a senior member of China Society of Technology Economics in 2004.

Yu You received his master degree in Software Engineering in October 2008 from Fudan University, China.



Wenjuan Wang received her bachelor degree in Information Security in July 2008 from Yunnan University, China. She started her master degree in System Analysis and Integration in September 2008, at School of Software, Yunnan University, China. She is current a joint student at School of Software, Fudan University, China. And her major research fields are e-business and mobile supply chain management. She has published three papers during the post-graduate.

Yiming Sun received his bachelor degree in Software Engineering in July 2009 from Fudan University, China. He started his master degree in the institute of E-business in September 2009, at School of Software, Fudan University, China. His interests include e-business model and techniques, business intelligence, data mining and citation network analysis. He has taken part in several research programs during the post-graduate.

Tong Li was born in Kunming, on December 24, 1963. He earned his Ph.D. in Software Engineering in February 2007 from De Montfort University, U.K, the B.Sc. degree in Computer Science in July 1983 and the M.Sc. degree in Computer Science in July 1988 from Yunnan University, Kunming, China.

He is now a professor and the dean of the School of Software at Yunnan University and the President of the Computer Society of Yunnan Province.

Prof. Li has published five monographs and over one hundred papers. His research interests include software engineering, concurrent processing and programming languages.