Design and Development of Semantic-based Search Engine Model

CAI Bo1，LI Yang-mei2

1.Wuhan Polytechnic, Wuhan, Hu Bei, 430074, China
2.Wuhan Vocational College of Software and Engineering, Wuhan, Hu Bei, 430205, China
583445738@qq.com

***Abstract*—With the rapid development of World Wide Web, search engines have become the main tool for people to get network information. However, the search results are widely criticized due to the poor accuracy and redundancy disadvantages. After the advent of semantic web, new search engine with the ability of understanding queries and documents has attracted more and more attentions. This paper starts from the traditional search engine, and firstly introduces its classification, popular technology, advantages and disadvantages, thus leads to the semantic search engine model. Then we research the semantic search technology in depth, which can be divided into enhanced semantic search based on traditional search, knowledge semantic search based on ontology inference and other semantic search types. In addition, key techniques in semantic search engine development, such as automated inference technique, ontology knowledge system and expert system are presented in this paper. Lastly, we conclude the current research and forecast the prospect of future research.**

Keywords-Search Engine; Semantic Search Engine; Ontology Inference

I. INTRODUCTION

Search engine uses software or manual methods to collect certain information with some strategies, and accomplishes information extraction, organization and processing to provide users with network information retrieval system. The massive content and hash organization of the Internet information bring great challenges for obtaining the search results accurately. Network information retrieval tools, especially search engine came into being under this situation. It provides an unprecedented high-speed network information resources retrieval platform, which has been basic services for users to get required information on the Internet, and has evolved a lot with the network technology and architecture upgrade. The advent of search engine brings users from the information passive acceptance age to the information initiative achievement age[1].

Nowadays, search engine is not only network information retrieval tools, but also has formed a huge industry, such as Google, Yahoo!, Baidu, etc[2]. In the wave of Internet technology, search techniques have transformed three times. With the global consulting networks are developing in the direction of interpreted machine, search engine will undergo another great changes and trials.

In this paper, section 2 provides a brief overview of traditional search engine techniques, advantages and disadvantages. Section 3 introduces the semantic search definition and conceptual model. Section 4 details the key techniques in semantic search engine, and section 5 concludes this paper.

II. TRADITIONAL SEARCH ENGINE

Traditional search engines are relative to the semantic search engine and intelligent search engines. Most of the search engine that we use today is traditional search engine. It is mainly for ordinary users and rarely uses semantic technology. Traditional search engine is established on the basis of grammar understanding of words and inadequate understanding of semantics. It regularly collects and organizes network information and provides keyword or browse query methods.

Traditional search engine has experienced a long period of development, and share diverse forms. Traditional search engine can be divided into two types according to the search content and search methods, and these two division methods can completely summarize the traditional search engine from two aspects: technology and form[1].

- According to the search content, traditional search engine can be divided into: search index/directory, full text search engine and meta search engine.
- According to the search methods, traditional search engine can be divided into: all-in-one search page, portal search engine, free for all links and vertical search engine.

Traditional search engine usually uses web crawlers to crawl information from webs. It extracts the keywords in the web with some strategies, and uses the keywords index. When a user submits a query to a search engine, it retrieves pages in the database satisfying the user's query requests with keywords matching approaches and feedback the results to the user. Key techniques in this process include word segmentation techniques, web crawler techniques, indexing techniques and word frequency index techniques. Word segmentation is a unique technique in Chinese semantic processing, and one of the most important indexes to evaluate the quality of Chinese search engine is the word segmentation technique[3].

The main advantage of traditional search engine is the ease of use. This search engine query method is simple and direct, response fast and searches for wide resource types, thus its retrieval efficiency is very high.

The limitations of traditional search engine are mainly embodied in the low recall ratio, low precision ratio, rough search results, weak multiple retrieval and poor understanding ability of the nature language.

978-1-4799-6636-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICICTA.2014.43

In order to overcome the disadvantages of the traditional search engine, semantic search engine has been proposed. It is an optimization and improvement for traditional search engine combined with advantages of other techniques.

III. SEMANTIC SEARCH ENGINE MODEL

Semantic search engine is available to analysis information object and retrieval request from the perspective of semantic understanding. It takes full advantage of semantic to overcome the disadvantage of traditional search engine.

The biggest difference between semantic search engine and traditional search engine is the intelligent search and content processing, therefore semantic search engine adds knowledge base and inference engine, forming a unique structure[4]. The structure of semantic engine is composed of crawler module, query engine, classifier, knowledge base and inference engine, as shown in figure 1[1].

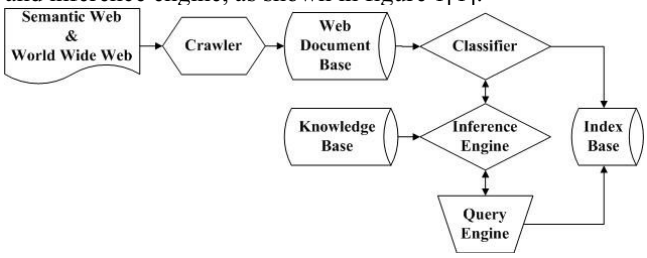

Figure 1. Semantic Search Engine Structure

- Crawler Module: Webpage browsing program, which can be ordinary crawler, semantic web crawler or a conjunction of the above two. It can also be "super crawler" which can browse all the web documents. The extracted data is stored in the search engine database.
- Query Engine: Analysis and processing of user queries, feedback the results matching the queries after semantic sort algorithm.
- Classifier: Process the data after inference engine and knowledge base, and establish indexes.
- Knowledge Base: It is composed of domain ontology or popularity ontology, which contains extensive classes and attributes examples, provides the knowledge base and inference rules in or across the domain, supporting for the semantic matching.
- Inference Engine: Analysis and inference of the user query and crawler extracted web content based on knowledge base.

According to the role of ontology technique, semantic search can be divided into three types: enhanced semantic search based on traditional search, knowledge semantic search based on ontology inference and other semantic search types.

A. Enhanced Semantic Search based on Traditional Search

Enhanced semantic search based on traditional search improves the traditional search results with semantic technique, the core of which is still traditional search engine. Ontology technique is used to enhance keywords search in a variety of ways to improve the recall and precision ratio of the search.

Reference [5] uses wordnet ontology for query expansion and defines synonyms. Keywords are retrieved in the ontology, and other different concepts are also retrieved by graph search. Words associated with these concepts are used to extend or restrict the search.

Tap is developed by Stanford University and IBM, which applies semantic Web technique to Google, to improve the search performance by increasing existing search results[6]. Match keywords with the concepts in RDF library, and the concept after matching is used to retrieve more files. In fact, it mainly gives answers to user query through traditional search engine. Only when the user query and entity in knowledge base are matched, structured semantic information associated with the entity will also be returned to users, and two different search results are displayed in one webpage. This method only respond to user's keyword request, and do not provide formal query capabilities, thus it lacks the ability of integration of two methods: information query and formal query.

B. Knowledge Semantic Search based on Ontology Inference

This type of semantic search is based on the ontology knowledge, and achieves knowledge semantic search by ontology inference. According to the different search objects, it can be divided into concept search and association search.

(1) Concept Search

According to the complexity of the user query, concept search can be divided into two types: simple concept search and complexconstrained concept search.

- Simple Concept Search: Knowledge base contains formal semantic information, mainly refers to concepts, intances and relationships. In fact users are not interested in abstracted domain knowledge such as the general concept of "human", but they focus on the specific instances information belonging to some concept. Therefore, the task of concept search is to retrieve all the instances belonging to a concept quickly and efficiently. General approaches used in simple concept search are match algorithm between keywords and concepts.
- Complex Constrained Concept Search: Some queries are complex, such as, retrieve all the teachers who were born in the 1960s and guild graduate students born in the 1980s to complete their postgraduate study. This is a typical complex constrained concept search. It usually needs to retrieve a set of concepts during implementation, and these concepts are connected together by some specific connections. In the semantic web environment, formalized the complex constrainted concept search is not difficult to achieve. The main difficulty for users is how to express the complex constrained query. Therefore, researchers have begun in the study of user query interface, aiming to provide a more intuitive query mode.

(2) Association Search

Most researchers difine semantic search to be search resources in the semantic web with basic inference function to find relevant resources needed by users. This definition satisfies user requests in most cases, but misses a class of actually existing search, which is called as "association search". Association search discovers and sorts complex relations between resouces in the semantic web, and object attributes are used to define associated relations between resources. Link path between resources may be more important than some certain resources, such as the links in national security may mean some potential security threats.

The main problem in association search is how to define the interest weight on the link, which can not only eliminate the associated links that users are not interested in, but also find the complex and hidden associated links in the data. Reference [7] proposes a popular and simple formal calculation method for discovering valuable relations between resources.

C. Knowledge Semantic Search based on Ontology Inference

University of Washington Turing Center is a multi-disciplinary research center. KnowItAll system developed by them is used to extract information on the web, aiming to construct an artificial intelligence system. The system can build knowledge library based on the whole web in a long life circle in order to replace search engine with information extracted way in the future. Reference [8] proposes several information extracted ways for searching useful information on the web and build domain knowledge library.

IV. Key Techniques in Semantic Search Engine Development

Semantic search engine is developed from the traditional search engine, which mainly relies on artificial intelligence techniques, automated inference techniques, expert systems, knowledge library system to optimize and update the search methods and search techniques. It changes a lot with the time to comply with the requirements of intelligence, personality and socialization[3]. In this section, we introduce three most important techniques as follows.

A. Automated Inference Technique

Inference is a form of logical thinking form of deducing new judgements or conclusions with known conditions. It is a commonly used method for problem solving which relies on people's mastery of knowledge and connections between things. Automated inference uses computer to imitate people inference process and get the answers. It is composed of program inference, program coccectness proving and expert system.

Program inference means computer algorithm. Programmers design algorithms to imitate people inference process and develop them with machine language. Thus computers can make automated inference with some premises.

Program correctness proving is formalize the process of proving theorems with some programs or algorithms, making computers automated prove the theorems. This is important to improve the precision of the inference result and prevent faulty inferences.

Expert system is the control and judgment system of the inference. Knowledge and logic stored in the expert system makes inference more intelligent, and results more reasonable and accurate.

B. Ontology Knowledge System

Ontology itself is a philosophical concept, and has been introduced to the computer science to represent the different definitions and their relations accurately and formally, which provides semantic relations between definitions. There are definitions of ontology, the most popular one is "ontology is an explicit specification of conceptualization." Conceptualized primarily means abstract description of objective things, the basic meaning is independent with the external environment. Clarification requires that concepts must be defined precisely, and represented standardized. Knowledge in ontology is commonly reconized, which can represent definations and their relations. Ontology can help understand user query from semantics, eliminate semantic phenomen and thus reflecting the real information requirement of the users.

Since ontology can effectively express and query knowledge, eliminate the semantic ambiguity between synonyms, support semantic discovery and automated semantic matching and combination, thus it can be used to build knowledge systems. Ontology knowledge system can improve the utilization depth of knowledge, helpful for hidden knowledge discovery and acquisition, knowledge sharing and innovation. The functions of ontology knowledge system are as follows: to identify a variety of language representation forms and storage forms; capable for ontology learning, ontology mapping, automatic ontology merging and other related operations; able to support the scalability and consistency of the ontology and compatible management of the ontology version.

As ontology knowledge system has strong semantic understanding and automatic inference ability, its processing and utilization of information is beyond the database system. Also, it can manage and manipulate information from the perspective of knowledge to make information retrieval rise to knowledge retrieval. Therefore, ontology knowledge system is an important technique in semantic search engines.

C. Expert System

Expert system is a kind of intelligent computer program system, which focues on the knowledge and experience of experts. It imitates the methods of experts to solve problems with computer programs to make the computer achieve the same level with the expert in problem solving, thus improve the intelligence of the computer. Thus, to build an expert system, a large number of expert knowledge is needed, and the programs must try to imitate the expert thinking ways to get problem solving skills.

As an intelligent system, it is very complicated to build an expert system. It is like a huge project, thus expert system is also called as “knowledge project”. Expert system

is usually composed of human-computer interface, knowledge acquisition, knowledge library, inference machine, interpreter machine and the comprehensive database.

V. CONCLUSIONS

This paper introduces the semantic search engine model and its key techniques. It has improved a lot from the traditional search engine, but still has some disadvantages, such as long delay of search operation, low scanning efficiency and un-mature semantic annotations.

Future researches on semantic search engine can be concluded as follows: (1) Research of semantic search concept model; (2) Construction, maintenance and evolution of semantic search ontology knowledge library; (3) Semantic search inference mechanism; (4) Sorting method of semantic search results; (5) Prototype implementation and application of semantic search engine.

REFERENCES

[1] Yiping Z. Comparative Study on Common and Semantic Search Engine[D].Jilin University, 2009.

[2] Kunmei W, Zhengding L, Xiaolin S, Ruixuan L. Research of Semantic Search: Overview[J]. Computer Science, 2008, 35(5): 1-4.

[3] Qian Z. A Comparative Study of the Traditional Search Engine and Intelligent Search Engine[D]. Zhengzhou University, 2012.

[4] Manjula D, Geetha T.V. Semantic Search Engine Journal of Information & Knowledge Management, Vol. 3, No. 1 (2004): 107-117.

[5] Moldovan D I, Mihalcea R. Using wordnet and lexical operators to improve internet searches[J]. IEEE Internet Computing, 2000, 4(1): 34-43.

[6] Guha R, McCool R. The tap knowledge base. http://tap.stanford.edu/

[7] Anyanwu K, Sheth A. P-Queries: enabling querying for semantic associations on the semantic web[C]//Proceedings of the 12th international conference on World Wide Web. ACM, 2003: 690-699.

[8] Popescu A M, Etzioni O. Extracting product features and opinions from reviews[M]//Natural language processing and text mining. Springer London, 2007: 9-28.