Research on Information Retrieval System Based on Semantic Web and Multi-Agent

LUO Junwei
College of Computer Science and Technology
Henan Polytechnic University
JiaoZuo, China
ljwonly@yahoo.com.cn

XUE Xiao
College of Computer Science and Technology
Henan Polytechnic University
JiaoZuo, China
50697895@qq.com

***Abstract*—The Semantic Web and Multi-Agent are effective means for constructing information retrieval systems. Despite a great deal of research, a number of challenges still exist before making Semantic Web and agent-based computing a widely accepted in information retrieval practice. In order to solve the problem of "difficult to feedback useful information to users", the paper developed a new information retrieval system which integrating Semantic Web with Multi-Agent to retrieval relevant documents or information by analyzing semantics contained in the queries and documents.**

Keywords-Semantic Web; Information Retrieval; Multi-Agent; Intelligent System; Data Mining

I. INTRODUCTION

The World Wide Web appeared around 1994. Its openness and broad accessibility greatly promote the development of information technology and its applications. Nowadays, various forms of digital contents like documents and web pages have been growing exponentially. In face of the overwhelming information volumes, people are struggling with information overload rather than its shortage. In order to handle multitudinous digital contents, information retrieval and related theories and technologies for the acquisition, management, and application of digital contents have risen as an important issue. But, it is well known that existing information retrieval systems based entirely on keywords have serious limitations and has led to the following problems: (1) Semantics in users' queries and documents cannot be extracted based on keywords. Most of the existing retrieval systems use keywords search and directory search. And the users' queries and the information in internet identified through keywords did not always meet the users' requirements. Because of synonym and polysemy in human language, information retrieval through keywords always left out other information with similar semantics; (2) Traditional retrieval system is lacking interaction with the users. According to some researches, the information of users' behavior can improve the rate of retrieval precision and retrieval recall. Even though some search engines record a large number of user behavior information through the log files, but they do not effectively use the information to establish the feedback mechanism to guide information retrieval and communicate with users. (3) Poor sort out search results. The results most search engines return are lack of precision. The users' search behaviors often bring a lot of spam. According to the evaluation of experts, the rate of relevant results the current major search engines return is less than 45%. At the same time, because the search-matching algorithms are not ideal, so the search engines sort search results too rough. These deficiencies have restricted the development of information retrieval[1,2,3].

Based on the above analyses, this study developed a new information retrieval system integrating Semantic Web with Multi-agent that handles the processing, recognition, extraction, extensions and matching of content semantics to achieve the following objectives: (1) Using Resource Description Framework (RDF) to analyze and determine the semantic features of users' queries, to present a new algorithm to extract semantics in the content and build up semantic database; (2) to present a new matching algorithm using semantics extracted from content which can feedback useful and accurate information meeting users' requirements; (3) a new Information Retrieval based on Multi-agent is put forward, the Agents in this model can adapt users' own interests and hobbies, collect information based on users' behavior, dig up semantics in internet and feedback and share information between different users, so the search results will be more in line with users' needs and help users to complete complex tasks.

II. INFORMATION RETRIEVAL SYSTEM MODEL

Agent technology appeared in the 70s of the 20th century. Agent can be defined as: an entity has the capacity of calculating, perceiving, reasoning, operating and communicating. Multi-agent system is composed of some Agents. Multi-agent technology can be applied to the research of information retrieval. The combination of information retrieval and Multi-agent technology has the following features: (1) Adaptability: Based on the information of users' behaviors in internet, Agent can discover the users' interest, reason the user's needs and establish personalized documentation for each user; (2) Initiative: Agent can initiatively retrieve the corresponding information based on users' demand, and even can monitor the changes of information sources; (3) Collaborative: Agents can share the information with other Agents. For example, a user's Agent can access to a lot of useful information from other users' Agents that have the same data about users' interest[4,5,6].

978-0-7695-4014-6/10 $26.00 © 2010 IEEE
DOI 10.1109/ICICCI.2010.35

IEEE computer society

RDF is a way for data and Meta-data representation. An RDF statement consists of triples (Subject, Property, and Object). RDF data model is a directed graph whose nodes are the subjects and objects and whose arcs are the properties. Nodes are labeled by means of URIs describing resources or literals (i.e. strings or numbers) or are unlabeled, called blank nodes. Blank nodes are usually used to group properties. Edges are always labeled by URIs representing a relationship between the subject and the object.

The information retrieval system designed utilizes RDF and Multi-agent technology as the basis to transform users' queries and documents in database into semantic pattern as triples: Subject, Property and Object, so as to process semantic, recognize semantic, and match semantic. Users can perform semantic-based and Multi-agent-based information retrieval in the following process:

First of all, users should submit their queries to User Agent, User Agent analyzed and determined user's characters about retrieval, and the query record will be stored in User Database, and the query will be transmitted to Extract Agent. Extract Agent will extract the semantic patterns in queries which can represent actual users' requirement. The next step was finished by Semantic Matching Agent, it will complete matching user's semantics with document semantic which stored in document semantic database, and the feedback the result to User Agent. User Agent will display search results based on user's character information in User Database.

Based on the users' different requirements, Information Gathering Agent can select different types of web robot to collect information in the internet and monitor the robot. The documents gathered by Information Gathering Agent will be analyzed and extracted semantics by Semantic Extract Agent, and the semantics in the document will be stored in Document Database.

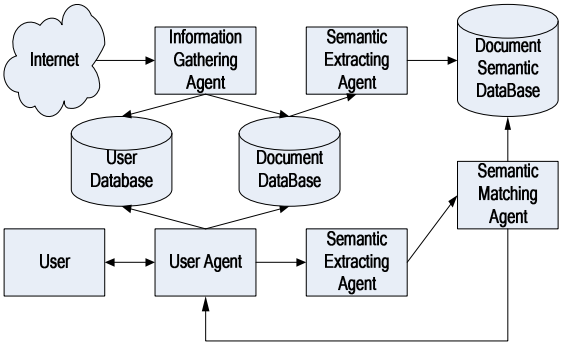

Figure 1. Information Retrieval System Model.

III. ARCHITECTURE OF INFORMATION RETRIEVAL SYSTEM

A. User Agent

User Agent includes Environment Perception, Memory Base, Knowledge Base, Learning Machine and Inference Engine. (Shown in Figure 2)

1) The perception of the environment module in the User Agent is the user's input and output interface.

2) Memory Base records the original information entered by the user.

3) Knowledge Base defines a user's personal knowledge, classified information and the user model.

4) Learning Machine is used to summary the behavior of users and formats the information.

5) Inference engine infers the user's interest based on the analysis.

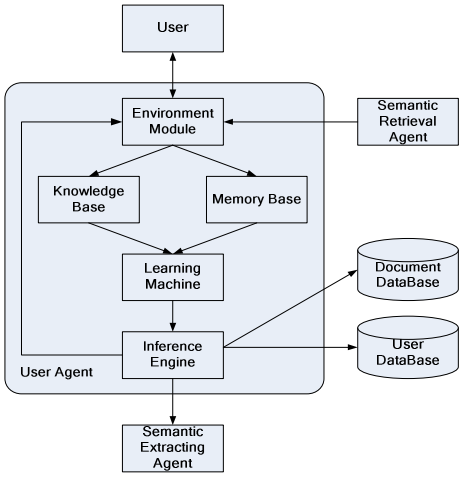

Figure 2. User Agent.

B. Information Gathering Agent

Information Gathering Agent mainly includes Search Strategy, Search Optimization and Robot. (Shown in Figure 3)

1) Search Strategy includes breadth-first search strategy, depth-first strategy and IP address search strategy.

2) Search Optimization includes the way accessing to the page should be subject to management of websites and the frequency of visiting, collecting important web pages which have high page weight and have changed, ensuring pages that will not be repeated crawled.

3) Robot is a program that crawl pages based on the hyperlinks between webs to collect information.

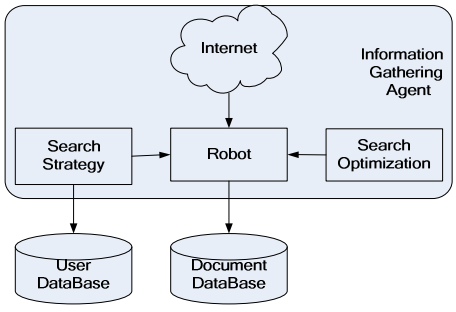

Figure 3. Information Gathering Agent.

C. Semantic Extraction Agent

Semantic Extraction Agent: it aims to mine the semantic features in the users' queries and documents. It will make

use of agent technologies, natural language processing technology, ontology technologies and other technologies to analyze the nature of words, structure and association relation in the users' queries and document to extract semantic features. This module contains the following components:

1) Content preprocessing: Meaningless words like neuter pronouns, articles, and symbols in the content will be removed. While processing documents, this component will filter the documents in different formats, as well as images, audio, video and other information formats, and identify and eliminate the noise content. Documents classification and index document will be also completed.

2) Semantic Analyzing: This component identified semantics elements like Subject, Property, and Object in the content and analyzes their semantic relations.

3) Semantic Extension: This component utilizes latent semantic analysis to create more semantic features for matching. This component solves the problem of query failure due to lack of required keywords in the query sentence even when only limited information of query content is available.

4) Semantic Construction: This component finally will construct the semantic pattern of the content according to Semantic Analyzing and Semantic Extension.

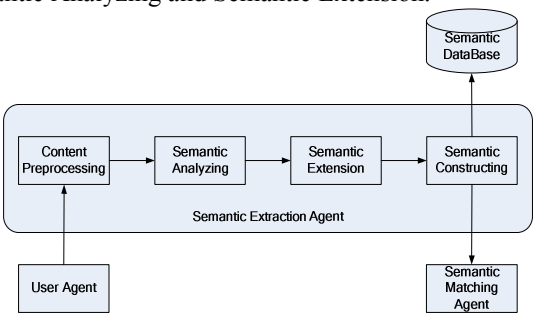

Figure 4. Semantic Extraction Agent.

D. *Semantic Retrieval Agent*

In the process, matching algorithms are presented to enable fast matching and searching for content. Components included in this module are: Knowledge Base, Model Base, and Semantic Matching. (Shown in Figure 5)

1) Knowledge base: It includes the user's personalized information transmitted by the User Agent. When matching, Semantic Matching will make use of uses' personalize information (users' gender, users' age, users' occupation, users' search history and so on) to match and search more accurate and useful information for users.

2) Model Base: It includes a variety of information retrieval model and matching algorithm, for example: Bayesian Probability Model, Support Vector Machine (SVM), Neural Network Algorithm and so on. It also includes new algorithms based on semantic pattern which can solve some traditional problems.

3) Semantic Matching: According to the Model Base, this component will chose apposite model and algorithm to matches semantics in users' queries and semantics in the documents, and in accordance with the relevant, the results will be submitted to the User Agent.

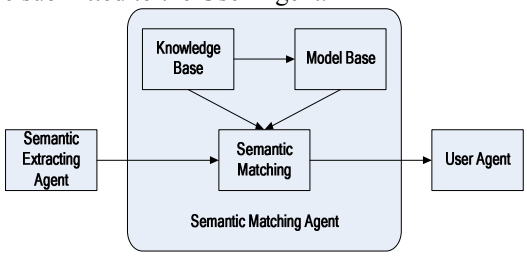

Figure 5. Semantic Retrieval Agent.

IV. CONCLUSION

In this study, a new information retrieval system based on Semantic Web and Multi-Agent has been presented to effectively offset existing defects and constraints of the traditional keyword-based search, and help users to obtain required information. This information retrieval system can be used in knowledge management, document management, search engine and other applications that require searching large quantities of information to achieve the purpose of reusing and sharing information.

REFERENCES

[1] SHI Xuelin, NIU Zhendong, SONG Hantao, "Intelligent Agent-based System for Digital Libraries Information Retrieval," Journal of Beijing Institute of Technology vol. 12, Aug. 2003, pp.450-454.

[2] J Naomi Aug, Augar, "Intelligent Information Agents: Search Engines of the Future," Proceedings of the First Australian Undergraduate Students' Computing Conference, Secec Press, Dec. 2003, pp. 10-16.

[3] Monica Rogati, Yiming Yang, "High-Performing Feature Selection for Text Classification," Communications of the ACM, vol. 9, May. 2002, pp. 4-9.

[4] P. Maes, "Agetns That Reduce Work and Information Overload," Communications of the ACM, vol. 37, Sep. 1994, pp. 31-40.

[5] Claudia Roda, Albert Angehm, Thieny Nabeth, "Using Conversational Agents to Support the Adoption of Knowledge Sharing Practices," Internet with Computer, vol.15, July. 2003, pp. 57-89

[6] D.Mladenic, "Personal Web Watcher: Design and Implementation. Technical Report," Dept of Intelligent Systems, vol. 21, March. 2005, pp. 12-20.