2011 International Conference on Information Management and Engineering (ICIME 2011)
IPCSIT vol. 52 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V52.78

Intelligent Information Retrieval Based on Ontology and Data Mining

Xiaohui Yu [+] and Xinsheng Ke

School of Economics and Management Beijing Jiaotong University, Beijing, China

Abstract. With the development of network information, intelligent information retrieval is the development direction of the search engine. Ontology and data mining may be the technology through which intelligent information retrieval technology can be achieved. By discussing the advantages and disadvantages of these two technologies, the author analyzes how to combine the advantages of two technologies to compensate for their shortcomings. To achieve better intelligence is the focus of this paper.

Keywords: ontology, data mining, intelligent information retrieval, complementary advantage.

1. The Development and Problems of Network Information Retrieva

1.1. The development of network information

In May 1998, Joint United Nations Information Committee named the Internet as the fourth media, marking the fourth media gain general acceptance of Information Media. Traditionally, newspapers, magazines and other print media is the first media, broadcast media the second, television media the third. Compared with the previous three, the biggest advantage of the Internet media is its speediness, interaction, and low cost. People can share the world's giant digital multimedia resource based on this platform, get the information they need, send and receive mail, download the necessary software and conduct online transactions.

Chinese access to the Internet was in 1994. Though later than others, the development is very fast. The "26th China Internet Development Report" of China Internet Network Information Center (CNNIC), July 15, 2010, shows that the scale of Chinese Internet users reached 420 million, and the scale of mobile phone users reached 277 million.

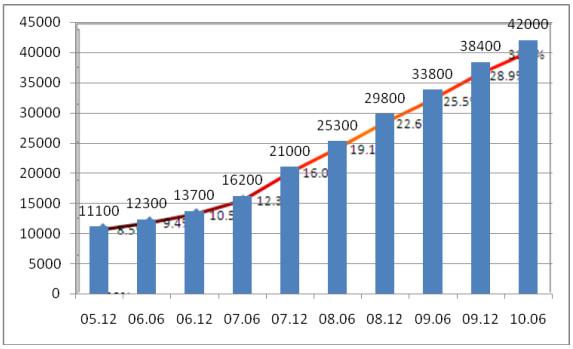

Fig. 1: the scale of Chinese Netizens and the popularizing rate

[+] Corresponding author. Tel.: + 010—51465258.
E-mail address: 10120754@bjtu.edu.cn.

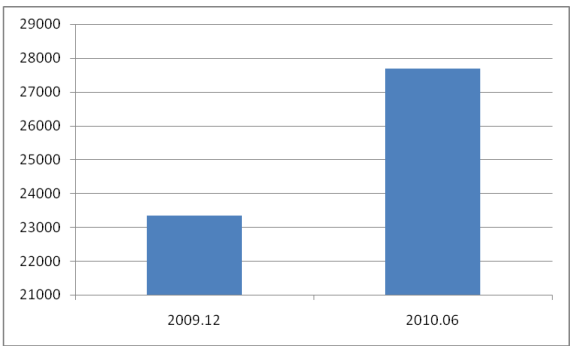

Fig. 2: mobile phone internet users size and growth conditions

According to the report, up to June 2010, the number of Chinese Netizens had an increase of over 36 million people than the end of 2009. Internet penetration rose to 31.8%, with 2.9 percentage points increasing compared with late 2009. The scale of broadband users was 363.81 million. Within the people access to the Internet, the penetration of Internet broadband reached 98.1%. The scale of rural Internet users reached 115.08 million, accounting for 27.4% of Internet users and six months increasing 7.7%, lower than the increase in urban users. The scale of mobile phone users in China reached 277 million, new mobile phone users increasing 43.34 million half a year, a growth of 18.6%. The scale of search engine users reached 3.2 billion people. Persons only using the mobile phone to surf the Internet took up 11.7% of the total Netizens.

In the process of using the Internet, the application behavior of getting information has been the main, which has a higher usage. The usage of search engine in 2009 was 73.3%, and to June 2010, this data increased to 78.3%, adding 5 percentage points. In terms of surfing the Internet with cell phone, mobile search ranks the second in behaviors of the application with a 48.4% usage.

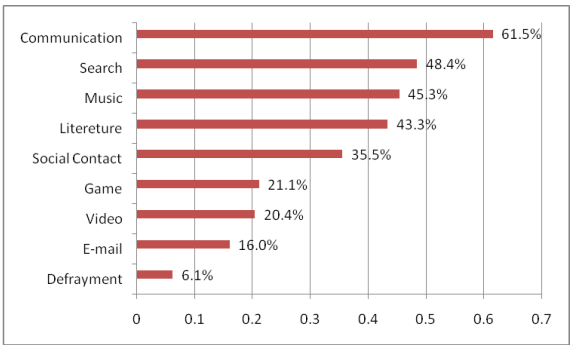

Fig. 3: web application of mobile phone users

1.2. The development and shortages of search engine

Since network information is rich and complex, the user must address the different needs of different circumstances with appropriate ways of retrieval, so that they can find the information quickly and accurately. Search engine is a tool which can retrieve information from a variety of network resource. Database information coming from the Internet websites is very large, and its retrieval speed is high. What's more, it can combine multiple keywords to improve search hits. But the search engine still has its own limitations. Web search engines are now mainly based on simple classification of information subjects or keywords, with problems in both entirety and accuracy. On one hand， the retrieval uses natural languages with inconsistency or local differences, which prone to the phenomenon of synonyms; on the other hand, facing massive retrieval results, it is still difficult for the users to get the desired ones; third, with a further growth of the World Wide Web, search engines will have to maintain the quality of its search results, but they only visit the static pages of the WEB ignoring the parts of dynamic pages generated by the database.[1]

Intelligence is the major development direction of information retrieval. Intelligent retrieval is a searching form based on natural language. The machine analyzes user's search requirement which is stated in natural language, and then comes up with a search strategy to search. In recent years, Internet keeps on emerging artificial intelligence products, such as Intelligent Search Engine, Intelligent Browser, Intelligent Agent,

Knowledge-Sharing Agent and so on. In general speaking, the intelligent search engine has three main features: web spider intelligent, provide relevant information for specific users, intelligent human-machine interface of search engines. It can guide the user in the Internet and expect the needs of users. It also can be used in information gathering, indexing, filtering (including interest filtering and poor information filtering), and through providing the user interesting and useful information automatically, it can provide personalized service.[2]

In view of the rapid development of network information mentioned above, problems that exist in search engine, and the development of intelligent search engine tools, exploring in network information retrieval theory becomes a research focus of information management. The intelligent implementation can be achieved with two technologies: ontology engineering theory and data mining technology. The former is used to achieve the accuracy of information retrieval, and the latter to implement a comprehensive information retrieval.

2. Way to Achieve Intelligent Search Engine

2.1. Ontology-Based intelligent information retrieval

Ontology is now widely used in research and application of artificial intelligence, computer science, information, becoming the latest research hot spots in intelligent network information retrieval. The basic principle is to establish some domain ontology with the participation of experts in this field; collect network information, and build knowledge database in accordance with the principles of ontology; converse the user search request to the ontological concept under rules and match in the Knowledge Database, then retrieve the results back to the seekers.[3]

The so-called intelligent search is that the user uses natural language to describe the retrieval request. The process of knowledge retrieval is the process to match the user's needs and network resources. Lack of experience as users and shortcoming of retrieval methods or other reasons, often it is difficult to return the real needs of the user's actual retrieval, but result a lot of useless information. In order to improve the situation, this paper optimizes and expands retrieval through various relationships in the ontology. Assume that the user's original need is {K, Q}, where K is the collection of the concepts of the user's retrieval, and Q is the logical relationships among the concepts. The process of using ontology languages to optimize the retrieval can be divided into two steps: the first step, map the concepts of K to the concepts of language ontology and form a new set of concepts C; the second step, according to semantic and the original logical relations Q, implement transformation rules on C, determine the new logical relationship among the new concepts, and form a new concept space. Semantic relationships among the new concepts include: synonymy relations, upper / lower relationships, semi-meaning/full-meaning relationship and antisense relationships. Apply the logic conversion rules to achieve conversion process. In order to accurately grasp the user's retrieval needs, return results of the user needs quickly, also need to build a users' preference ontology database.[4]

2.2. Intelligent search based on data mining

Data mining is to dig out the hidden knowledge from the large amount of data. The concept of data mining should be divided into two: special and general. as a general rule, general data mining is also known as Knowledge Discovery in Database (KDD). It is to extract of implicit, unknown in advance, but credible, the potential and the valuable information and knowledge from a large number of incomplete, noisy, fuzzy and random data. Special data mining is to use various data analysis tools to find model and the relationship among data from mass data, which is a process of KDD. Although the time that data mining came into being is short, it's a computer analysis technique with speed development. In these years, data mining is day by day mature, and shines in practice, and is applied in various fields. The most commonly used includes decision trees, neural networks, association rules, cluster analysis, statistical fuzzy sets and rough sets, etc.[5]

Data mining is to find the implicit regularity from the large amount of data, and its application is to resolve the problems of data quality. That data can be more efficient used and the useless be discarded, is the most important forms of realizing data mining and its main application. Data in traditional database compared with those on the WEB, its structure is very strong, and is completely structured, while semi-structured data is

the most prominent feature on the WEB. Thus, data mining in a single data warehouse is much easier than from the WEB.[6] According to statistics, a lot of contents the website provides for users are useless. The fact is that for a particular user, the content of interest is only a very small part of the site, and most contents available on the website for this user is useless, and often too much information may hide the useful one, making data mining more inefficient, and the site undesirable. To the WEB data mining technology, retrieval and integration of the semi-structured data model and source model should be the primary problem, which is the prerequisite for the WEB data mining to be accessed.

3. The Integration of Two Technologies

3.1. The application of data mining in ontology specifications

An important tool that ontology engineering uses to achieve intelligent information retrieval is to build personal preference ontology database. The database stores preferences and retrieval interests of user's, and updates and refines constantly according to retrieval of user's long-term retrieval. Ontology uses classification and clustering techniques to organize search results, user's browsing information of the results classified to build user's preference database as the integration results of original search engine. In another word, , monitor the user's WEB page browsing behavior, the time he or her spent on a page, the length of the document, the URL address to create log file. Then by analyzing the log file the system can get the user's interests, and model user's interests. One drawback of this approach is that the system can only passively accept the user's browsing behavior during the process of getting use's interests, which is a simple superposition of the data, without deeply analyzing the user's browsing behavior. Data mining technology can solve this problem perfectly.

Data mining can find hidden and regular contents from a lot of data, which can be applied to solve the problems of quality of data application. So that data can be used more efficiently and the useless data be discarded. Analyze the user's behavior, according to association rules, to optimize the organizations of website information and to provide users the pages maybe needed; through cluster analysis cluster users with similar access characteristics and pages with relevant content to provide users relevant hyperlink according to the user's requirement; through the classification rules, use statistical methods, machine learning, neural network to construct a classification model, draw the outline of characteristics of the user; through sequential patterns, predict the user's access patterns, and understand the user's interests and needs.

4. Conclusions

Today with more and more knowledge, more and more web content, as a new emerging field in the realization of the popular intelligent network technology, Ontology and data mining technology have been paid more and more attention, and in many ways come into widespread use. However, both are immature in theory and practice and have certain limitations, which make Ontology and data mining have limited application. Therefore how to combine the advantages of the two to provide human more intelligent information retrieval services needs further consideration.

5. References

[1] Chun Chang. The Problems and Development Direction of Internet Information Retrieval [J]. Journal of Information, 2003(11)

[2] Ruhua Huang, Chunlei Zhang. The Development Direction of Internet Information Retrieval [J]. Document ,Information and Knowledge, 2002.8

[3] Huadao Cheng. Ontology Application on Network Information Retrieval [J]. Journal of Hubei University of Police, 2010.3

[4] Xueqing Li, Yuwen Jia. Personalized Internet Information Retrieval Based on Ontology [J]. Researches in Library Science, 2007.1

[5] Wen Wang.Survey on Data Mining Technology [J]. Computer Knowledge and Technology, 2010.8

[6] Jinlong Ding.Personalized Information Service Based on Web Data Mining Techniques [J]. Journal of Modern Information,2010.3