

Research Proposal

Title of research: Knowledge Discovery from Web Search

Name of student: Tu-Liang Lin

Keywords: knowledge discovery, search engine, natural language processing

Knowledge discovery is concerned with an information extraction activity whose goal is to discover hidden and valuable knowledge from a certain domain. Using a combination of machine learning, statistical analysis, search engine, modeling techniques and natural language processing, knowledge discovery techniques find significant terms and relationship between these terms and moreover conclude some rules that help people to make decisions. Unfortunately, knowledge discovery system often outputs some useless rules and perplexes users because knowledge is hard to define. Some rules are maybe useful to some people but to other people these rules are garbage information. If people use tremendous data like Internet as input source, the problem will be more serious.

Because of the popularity of computers and networks, Internet has become the most important information source. Knowledge discovery from web search is a technique that extracts knowledge from Internet, using search engine, spider and natural language processing. Traditionally, people use some keywords and simple Boolean

algebra to find out the related articles, as figure (1). Although search by keywords is the most efficient and popular method to find related information in the Internet, it exists two problems by using this method. The first is that some search results don't match the user's requirement. The other is that there are too many similar articles in the search results. Because of the two problems, users spend a lot of time organizing the search results and finding what they really want.

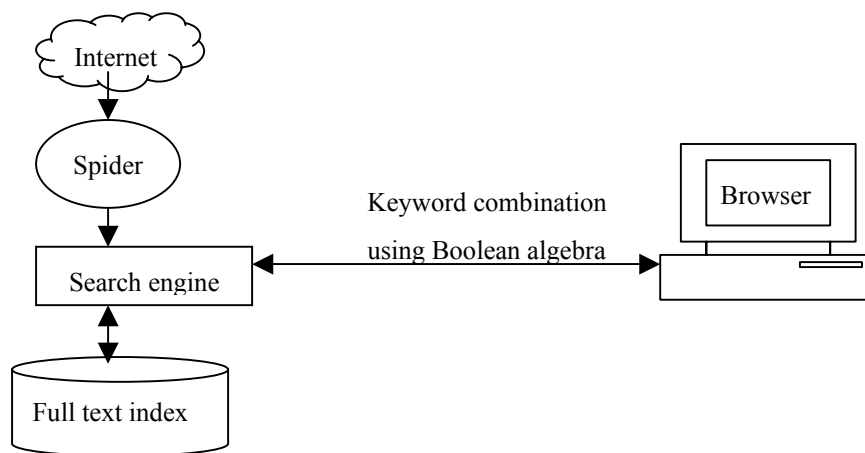


Figure (1) Search by keywords

Because of the imprecise results of keyword search in the Internet, all the studies of web mining method are trying to improve the accuracy or value of the information gotten from the web pages. According the research of Facca, F. M. and Lanzi, P. L. at 2003, web-mining methods can be categorized to three basic directions, web content mining, web structure mining and web usage mining. Web content mining focuses on the row information available in web pages and source data are mainly text. The most successful applications of web content mining are content-based categorization and

ranking of web pages, which are adopted by many search engine companies, such as google, altavista and lycos. Web structure mining focuses on the structure of web sites and source data are mainly the structural information, such as links to other pages. Web usage mining extracts information from sever log files. Web content mining is the most challenge domain among the three types of web mining, because of the variety of web content.

When search engine companies face web content mining, they rarely get into in-depth linguistic analysis of document collections because of the complexity and time-consuming process. Although automatic language processing are much more complex and slower than full index processing, many researches about automatic language processing have been done recently and web analysis and search field increasingly adopts automatic language processing in web applications. (Chakrabarti, S.,2003) Some answers of knowledge inquiries are hard to obtain from keyword searching method, such as salesmen want to compare property value of a specific region. Therefore, a combination usage of automatic language processing, search engine, knowledge-base system, etc is important in solving questions raised by human beings.

In this research, I hope I can make some improvements in search engine field by using combination technology of knowledge-based system and web-mining method. Knowledge-based systems mainly contain two modules, knowledge module and control module. The knowledge module is called knowledge base, built by facts and rules, and the control module is called inference engine. (Hopgood, A. A., 2001) If web-mining technology is used to extract rules and facts from web pages, because of the unreliable characteristic of the Internet data, some conflicts will appear between the extractive rules and facts. Using statistic model and statistic data that is generated from search engine can solve these conflicts. After knowledge base based on the Internet data and statistic data is generated, it can provide censorship mechanism for search results and play an important role in information quality control.

Except knowledge base system, I will introduce question analyzer and text analyzer, based on natural language processing techniques, in this research. These two analyzers, adopting semantic analysis by using lexical networks such as WordNet (Miller, G. et al., 1993) or LDOCE (Longman Dictionary of Contemporary English) (Dolan, W. et al., 1993), are responsible for parsing questions that are inputted by users and texts that are outputted by search engine, as figure (2). Question analyzer distinguishes the question types and generates keywords from questions. After keywords are generated, search engine uses keywords to inquire the full text index db

and outputs some related results. Text analyzer collects all these results and tries to answer the questions that user asked using semantic analysis with the help of knowledge base system. For example a question, such as Where is Tony Blair's hometown? , is entered into the intelligent system and I hope the system can base on the analysis to produce the answer Edinburgh.

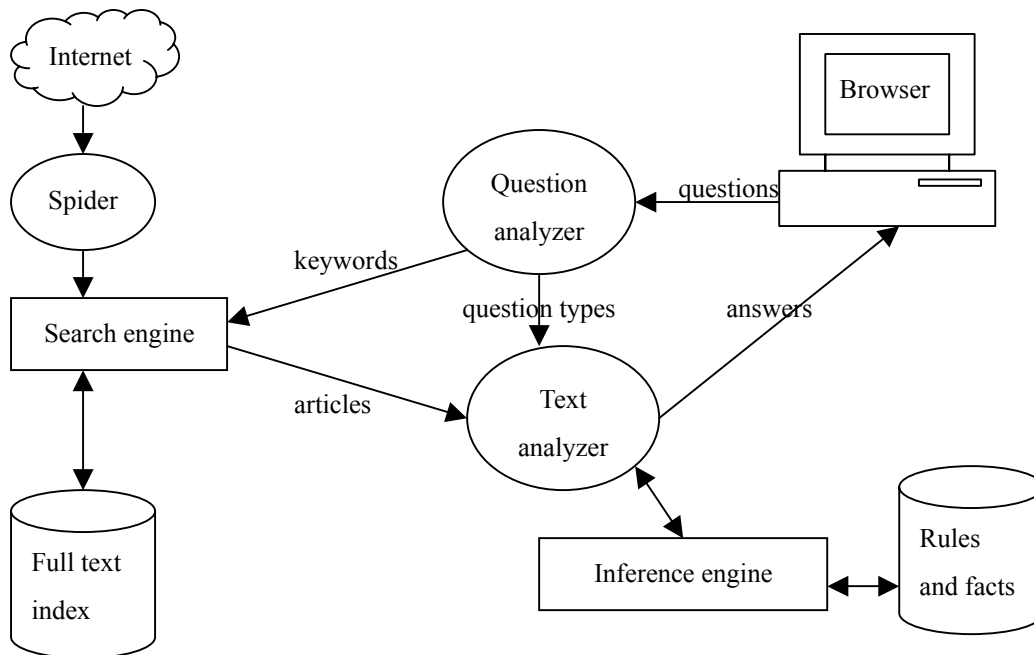


Figure (2) Intelligent web query system

This kind of QA (question answering) systems can be divided into three grades, slot-filling, limited-domain and open-domain, according their sophistication.

(Chakrabarti, S., 2003) Slot-filling QA system deals with special design questions.

Limited-domain QA system applies the application on specific domain. Open-domain

QA system is the most sophisticated one among the three grades because it deals with the questions without any restriction and consequently it is still an intensely research field. The intelligent web query system belongs to open-domain QA system because of the variety of web content.

Although the intelligent web query system may work properly in some cases, it still has two difficult problems that need to be conquered in the new system design. The first is that search engine may outputs a lot of articles, and text analyzer cannot process so many articles in a batch in a short time. The second issue is the precision of question analyzer and text analyzer and this issue depends mainly on the intelligent algorithm design. Of course, the spider's intelligence is also a very important part in the architecture. There are three parts that we can put artificial intelligence in, including spider, question analyzer and text analyzer. Therefore I hope that the whole research will emphasizes mainly on artificial intelligence, especially natural language processing.

Finally, I think the fundamental goal of this research is to make some contributions in web search field and create a more convenient knowledge platform in Internet. Of course, there are some similar approaches, such as START (<http://www.ai.mit.edu/projects/infolab>) and Ask Jeeves (<http://www.ask.com>).

However these approaches all face the two problems I mentioned above, and there are still some aspects that are worth researcher to make a great improvement.

Reference

Chakrabarti, S., 2003, *Mining the web: discovering knowledge from hypertext data*,

Morgan Kaufmann Publishers, San Francisco

Dolan, W. et al., 1993, *Automatically Deriving Structured Knowledge Bases From*

On-Line Dictionaries, <ftp://ftp.research.microsoft.com/pub/tr/tr-93-07.ps>, Microsoft

Facca, F. M. and Lanzi P. L., 2003, *Recent Developments in Web Usage Mining*

Research, In Kambayashi Y. et al., *Data Warehousing and Knowledge Discovery:*

5th International Conference, DaWak 2003 Prague, September 3-5, 2003

Proceedings, Czech Republic, page 140-150

Hopgood, A. A., 2001, *Intelligent Systems for Engineers and Scientists*, CRC Press,

Boca Raton

Miller, G. et al., 1993, *Introduction to WordNet: An On-line Lexical Database*,

<ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.pdf>, Princeton University/