Present Situation of Web Information Search and Its Improvement Strategy

Lin Tang[1,2,3, a *]

[1.] Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China

[2.] School of Math and Computer Science, Mianyang Teachers' College, Mianyang 621006, China

[3.] Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis, Guangxi University for Nationalities, Nanning 530006, China

[a] tang_linmail@sina.com

Keywords: Web information search; Keyword search; Ontology knowledge base; Networked automatic reasoning

Abstract. With the advent of the era of big data, people face Web information overload problem. In order to efficiently and accurately obtain information on the network, Web information search technology has been developed rapidly. First in this paper, the working principle of the keyword search engines that are the traditional Web information search technology is reviewed, and then from the recall ratio and precision, reasoning search, information in a timely manner and the user experience etc, we analyze its existence disadvantages and Improvement strategy. Finally we present the train of thought and key technologies of a new Web information search that is based on networked automatic reasoning. This paper provides a train of thought for building automatic reasoning system which is adapted to the network of open, distributional and dynamic, and also puts forward a new direction for the development of Web information search technology and the architecture of next generation of search engine.

Introduction

As the Internet's rapid emergence and the steady accumulation of network resources, Web data show explosive growth. Due to a sharp increase in online information， in life people get convenience at the same time also face the problem of "information overload". That is to say, Information resource is rich, but really useful information is relatively scarce. This problem led to the development of Web information search technology. Search engine is the outstanding achievement of modern Web information search technology. Representative of search engines, such as Google, Yahoo, Alta Vista and baidu has become an essential tool for every network user. Users can use search engines to obtain useful information from the vast data, and also can greatly save the query time and improve the query efficiency.

The traditional way of Web information search is based on keyword search, that is based on string matching. Because it is only with the help of character manipulation and probability statistics method based on keyword to the network information retrieval, it is not well meet the demand of users' search in terms of recall, precision and reasoning search, information timely and user experience feeling, etc . So this paper will first review the working principle of the search engine based on the keywords, and then analyze its main problems to explore the improvement of search technology strategy. Finally we present the train of thought and key technologies of a new Web information search that is based on networked automatic reasoning.

Principle Based on Keyword Search

The whole system is made up of crawlers ,crawl control module, page library ,collection analysis module ,retrieve library ,order module and query module etc. In the process of system working, various modules Collaborate closely and cooperate with each other. Its working process is roughly as follows[1,2], First of all, the crawlers Crawl from the Internet to retrieve the web page, and would be to send the URI extracting from the web page to the crawl control module. Crawl control module decides the operation of crawl, namely the decides the next access link, and sent the link back to the crawlers. Crawlers keep crawling the entire network, and at the same time send the retrieved web page to the page library, until all of the pages have been grabbed or local resources exhausted; And then the collection analysis module parse web pages to obtain text information, and then sent to retrieve library. At the same time, system can extract and analysis the web link information in the retrieve library and transmit the information to the order module, providing the basis for future web sorting; Finally when the user submits the retrieval requirements, the query module is responsible for receiving and explaining the user's search request. Sort module classifies the results, and orders these according to the evaluation , and extracts the contents of the keyword. the system will eventually deal with a good page results back to the user.

Disadvantages Based on Keyword Search

By the principle of work, based on the keyword search retrieval object is not the actual Internet network, but the object is the After processing the retrieval library; Retrieval is the process of mechanical character matching, that only takes the grammar level into consideration while ignores its semantic understanding and the hides relationship between information and information. So it causes the loss of retrieval of information, and its search results is still far from satisfactory. Then the deficiency of traditional search engines that based on keywords or text content based retrieval are discussed as follows[3-5].

1. Low precision and the results of the search is a single page. Roughly two aspects, one is even search to the relevant page, but at the same time, people will face tens of thousands of the condition of the mixed relevant or irrelevant information. Thus the search has the disadvantages as follows return the relevant information too much, high cost in terms of time for positioning information needed by the user , and need artificial selection, etc. Secondly the computer can't understand the real "semantic" users type in a keyword, so returned to the many pointless web pages.
2. The recall rate is low. The reason to roughly two aspects, one is because the data on the Internet almost every day exponentially, however search engine retrieval database page growth lags far behind the growth rate of network resources. Users can access to the Internet through search technology up to 70% of the resources; The second reason is the semantic search results of keyword sensitive enough, and lead to low recall rate. Because of lack of semantic understanding of keyword search engine, while merely matching keywords itself， Even on the Internet there is useful information is expressed by the keyword synonym, often also can't search.
3. Search lacks of reasoning . Search results is only satisfy matching exists in the information on the network, and can't draw implicit knowledge by reasoning. Users will be read by multiple web pages on the relevant information, combined with their own reasoning ability to get the answer. It is difficult to automatically find effectively, integrating information and knowledge on the Web by the search engine.
4. The information has expired. Information on the web is dynamic. In the world of the Internet, information data are in a constant process of dynamic updating, and have very strong timeliness. Even if in the process of search, information will be add, modify and delete. However, The renewal of the crawler is not real-time that cause the search results "overdue".

Improvement Strategy of Web Search

In view of the above analysis, based on keyword search engine has obviously can't satisfy people's demand for Web knowledge discovery. For this Web search has roughly the following improvement strategy.

Use of the Traditional Information Retrieval Technology Using the traditional IR technique to improve the problem of precision low and Information expiration. In view of the above the first question, the precision is low. The IR technology is general improving the precision by improving the sort algorithm of search results and against spam information method . In terms of improvement of sorting algorithm, literature [6,7] the PageRank algorithm was optimized , and the literature [8] proposed combined scheduling problems with distributed environment. Literature [9] analyzed the influencing factors of sorts, in certain cases focuses more on the single factor. Literature [10] from the Angle of the multivariate vector improved sorting algorithm in order to obtain better results returned; In terms of research against spam information, the literature [11] StandFord university classified garbage information, and put forward how to combat spam will one of research focuses in the future, and in the literature [12,13] Lehigh university spam detection method was put forward.

In view of the above the fourth question, information expired. The IR technology usually adopt parallel crawler method and improve update website scheme. Literature [14] realized crawler parallel; Literature [15] put forward the importance of update website maintenance, and literature [16] improved update website. Literature [2] the use of metadata cooperation to complete web site updates, and literature [17] a classification method was put forward by way of the update process to improve the degree of crawler update.

Combination of Traditional IR and Semantic Web Technology This kind of semantic search which combined with traditional IR and semantic Web technology uses Web Ontology Language (OWL) to describe all the information in the network, and obtains semantic hidden in information by the semantic reasoning technology . To improve the recall ,precision, and query efficiency. Semantic search improved traditional IR technology from two aspects, on the one hand is that it takes advantage of the result of the semantic reasoning to enhance the traditional IR based on keyword search. In literature [18,19] extended query by synonyms set defined by WordNet ontology, and literature [20] used the above method based on geographic information ontology in the field of geographic information system to improve the retrieval results. in [21,22] Stanford University and IBM developed the Tap system which applied semantic search technology to Google, and in literature [23] the corresponding knowledge bases of Tap system were introduced. Keyword matched the knowledge base in the library (RDF) concept, using the concept of matching the search more information; On the one hand is the use of semantic web technology to improve the search itself. In the literature[24] ontology retrieval model based on vector space model was put forward. The model needs to transfer keyword query into the structure, rather than the combination of structure and content retrieval. Literature [25], selected by the user in the ontology concept hierarchy tree to constrain the concept of search.

However, the semantic search object is still the traditional information resources, is not the semantic resource. Does not support the formal inquiry at the same time, and only response the keyword query. As a result, the method of search results only in the recall ratio and precision improvement, and in the aspects of implicit information in web search and information timeliness etc, there are still based on the inherent disadvantages of traditional information resources search.

Semantic Search Based on Ontology Currently, more and more Web information resources Semantic annotate by using semantic Web languages, such as RDF [26] and OWL [27], etc. This makes in the semantic Web, there is a lot of structured, machine understandable object information, and it provides the foundation for semantic search. Search object [28] of this method is mainly composed of domain ontology knowledge base, through the technology of semantic web, automatic reasoning of knowledge mining and knowledge discovery model of semantic search. Ideal semantic search system workflow is roughly as follows: firstly, the user's questions process by the user interface module. User interface module understands the problems of semantic by semantic technology such as grammar and lexical analysis, and formalizes the problems; Automatic reasoning technology is then used to get the required information in the domain ontology knowledge base; Finally, the results sorting processing module and returned to the user. This way of searching from the recall ratio and precision ratio, reasoning ability and information in a timely manner can improve the search results.

However, the current domestic and foreign research and development of semantic search is still in its primary stage, did not form a perfect architecture. Has formed some related system can only provide the formal inquiry, only through the knowledge base of ontology reasoning extension keywords, improve the recall and precision, only using semantic sort to improve results. In order to achieve the ideal first need to depict knowledge semantic search system, namely, build and maintain for semantic search domain knowledge base; Need to extend reasoning mechanism to adapt to the network needs; Need to study for semantic search results sorting method; In the end, the need to explore the semantic technology and automatic reasoning model for semantic search.

Build a New Type of Search Engines Based on Networked Automatic Reasoning

Construction Idea In view of the above to the traditional based on the analysis of existing problems of keyword search engine and its improvement strategy of exploration. Adopt the method of automatic reasoning in advance to the user to provide response to system returns is no longer a list of web pages, but direct answer [29,30].But because the information on the network exists in multiple access network nodes can be, with a space level of dispersion, and dynamic. In order to get satisfactory results reasoning we need multiple nodes information share. To achieve cross knowledge reasoning, solve the train of thought has the following two kinds.

1. Centralized fusion

Centralized fusion, distributed information fusion in each network node into a knowledge base, using the automatic reasoning mechanism to provide response. Such as University of Washington Turing center in order to form in a long enough period based on the knowledge base of the whole Web, developed KnowItAll system [31,32] used to extract information on the Web. However fusion knowledge base will bring many problems [33] : for example the result of a merger knowledge base is not independent that is knowledge repetition and waste of space, and inconsistent that is Combined the knowledge of contradiction and delay problems [34-36].

2. Keep the scattered, build networked automatic reasoning mechanism

Using the various nodes on the network of the knowledge base for automatic reasoning, this mechanism is applicable to Web distribution, open and dynamic. However, the traditional knowledge representation is centralized, and traditional logic reasoning system is based on "closed world assumption" [37]. In other words, the reasoning mechanism of existing only within a single knowledge base, and the multiple knowledge base on reasoning has some limitations. Therefore constructing networked automatic reasoning mechanism and the reasoning across the knowledge base system is a big challenge. The inference system can unite different reasoning of knowledge base, reasoning task together. To a greater degree to meet the demand of the user's retrieval, but also avoid merging problems brought about by the knowledge base.

Key Technology We will combine the semantic, automatic reasoning and theorem proving technology to build networked automatic reasoning system, and its application to the design of new type of search engine. Specific key technology research is as follows:

1. Research the model of networked automatic reasoning

Traditional search model is not completely suitable for semantic search. on the basis of the existing search model, combining with formal semantic information, we need study the semantic search model based on semantic Web technology. The model will improve the results of current search engine , and become the base of a new generation search engine on the semantic Web in the future.

2. Research for networked automatic reasoning knowledge description method

In the current fields to establish ontology knowledge base in the industry, and ontology is with semantic, can be Shared and reasoning of knowledge expression. So as a logic based ontology description logic (DL), used to depict knowledge become the preferred, and its powerful description ability has also been widely used. However, reasoning efficiency and the power of expression is the description logic system still problems to be solved.

3. Build networked automatic reasoning mechanism

Based on ontology knowledge retrieval source, setting up across the ontology knowledge base networked automatic reasoning mechanism, is the new direction of information retrieval. To build this system is roughly divided into two aspects, first of all should be established between the knowledge base of inference rules, and then prove that the networked automatic inference system basic properties, such as reliability and completeness , etc.

4. The new search engine prototype implementation and application research

Based on the above research foundation, the networked automatic inference system is applied to design a new search engine. It will achieve certain areas such as urban traffic response system prototype, and test in the application environment and realize the performance optimization, giving users a satisfactory response [38].

Summary

This paper is the traditional search engine is introduced and analyzed the working principle and the existing problems, and then summarizes the current Web information search of three types of improvement strategies. Finally this paper puts forward the idea of a new type of search engine based on networked automatic reasoning, and points out the idea and key technology. This is a pile of system engineering, with a comprehensive and challenging. This is combined with the frontier of computer science and information science, innovative and forward-looking, also has the very good application prospect and social and economic benefits.

Acknowledgements

This work was supported by the Scientific Research Project No. 15ZB0284 from Sichuan Education Department, and the Scientific Research Project No. HCIC201405 from the open fund of Guangxi Key laboratory of hybrid computation and IC design analysis.

References

[1] Arvind Arasu, Junghoo Cho, Hector Garcia-Molina, et al. Searching the web. ACM Transactions on Internet Technology, 2001, 1(1): 2～43.

[2] Wen Kun-mei and Lu Zheng-ding. A Cooperative Schema between Web Server and Search Engine for Improving Freshness of Web Repository. Wuhan University Journal of Natural Sciences，2006,11(1): 11～14.

[3] T. Berners-Lee, J. Hendler, O. Lassila. The Semantic Web. Scientific American, 2001, 284(5): 34-43.

[4] Weizhe Zhang, hongli Zhang, Xiao Xu, et al. A distributed search engine system performance modeling and evaluation. Journal of software, 2012,23(2):253-265.

[5] Jianying Xu. Development trend of search engine study. Modern intelligence, 2011,9(31):51-55.

[6] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University Database Group,1998. Available at http: //dbpubs.stanford.edu: 8090 /pub/1999-66 .

[7] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating pagerank computations. In Proceedings of the International World-Wide Web Conference. New York: ACM Press, 2003. 261～270 .

[8] Y. Wang and D. DeWitt. Computing pagerank in a distributed internet search system. In Proceedings of the 30th International Conference on Very Large Databases. 2004. 420～431.

[9] Kim, S.J., and Lee, S.H. An Improved Computation of the PageRank Algorithm. In: Crestani, F., Girolamo, M., and van Rijsbergen, C.J. Proceedings of the European Colloquium on Information Retrieval. Springer LNCS 2291, 2002: 73～85.

[10] Taher H.Haveliwala. Topic-sensitive PageRank: A contextsensitive ranking algorithm for Web search. IEEE Trans. Knowledge and Data Engineering, 2003, 15(4): 784～796.

[11] Zoltan Gyongyi and Hector Garcia-Molina. Web spam taxonomy. In Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005 .

[12] Baoning Wu, Brian D. Davison. Cloaking and Redirection: A Preliminary Study. In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), May 2005.

[13] Andras A. Benczur, Karoly Csalogany, Tamas Sarlos, et al. SpamRank – Fully Automatic Link Spam Detection. In Proceedings of the First International Workshop on. Adversarial Information Retrieval on the Web (AIRWeb), May 2005.

[14] J. Cho and H. Garcia-Molina. Parallel Crawlers. In Proceedings of the 11th International World Wide Web Conference. ACM Press, 2002. 124～135.

[15] Cho J, Garcia-Molina H. Synchronizing a database to improve updating. In: Proceedings of the International Conference on Management of Data. 2000. 256~262.

[16] Kunmei Wen, Zhengding Lu, Weiguo Ye, Li Jin. Analysis and improvement of the search engine page update strategy. Journal of huazhong university of science and technology (natural science edition),2002, 30(12):3~5.

[17] Kunmei Wen, Zhengding Lu. Search engine based on the classification of the web pages updating method research. Journal of computer science, 2004, 31(9A):1~2.

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

[19] Kruse, P.M., Naujoks, A., Roesner, D., Kunze, M.Clever search: A wordnet based wrapper for internet search engines. In: Proceedings of the 2nd GermaNet Workshop, 2005. 367 - 380.

[20] Buscaldi, D., Rosso, P., Arnal, E.S.A wordnet-based query expansion method for geographical information retrieval. In: Working Notes for the CLEF Workshop, 2005.

[21] Guha, R., McCool, R.TAP: A Semantic Web Test-bed. Journal of Web Semantics, 2003, 1(1): 81~87.

[22] R.Guha and R. McCool. Tap: Towards a web of data. http://tap.stanford.edu/ .

[23] R.Guha and R. McCool. The tap knowledge base. http://tap.stanford.edu/.

[24] Vallet D, Fernández M , Castells P. An ontology-based information retrieval model. In Proceedings of the 2nd European Semantic Web Conference (ESWC). New York:Springer, 2005. 455-470.

[25] Airio, E., J¨arvelin, K., Saatsi, P., Kek¨al¨ainen, J., Suomela, S. Ciri - an ontology-based query interface for text retrieval. In Proceedings of the 11th Finnish Artificial Intelligence Conference. 2004. 73~82.

[26] Manola F, Miller E. Rdf primer. W3C Recommendation, February 2004. Available at http://www.w3.org/ TR/2004/REC-rdf-primer220040210/.

[27] McGuinness Deborah L, Harmelen van F. Owl web ontology language overview. W3C Recommendation, February 2004. Available at http://ww.w3.org/TR/owl-features/.

[28] S. Staab, J. Angele, S. Decker, M. Erdmann, A. Hotho, A. Maedche, R. Studer, and Y. Sure. Semantic community web portals. Computer Networks, 2000,33(1-6):473-491.

[29] A .M .Turing. Computing Machinery and Intelligence. Mind, 1950; 59 (23 6):433-460

[30] R. B. Mishra and S. Kumar. Semantic Web Reasoners and Languages. Artificial Intelligence Review, 2011, 35(4): 339–368.

[31] D. Downey, O. Etzioni, and S. Soderland. A Probabilistic Model of Redundancy in Information Extraction. In Proceedings of the 19th International Joint Conference on Artificial Intelligence, 2005. 1034~1041.

[32] Michael Cafarella, Doug Downey, Stephen Soderland, and Oren Etzioni. KnowItAll: Fast, Scalable Information Extraction from the Web. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2005. 563~570.

[33] Gregoire E, Konieczny S. Logic-based approaches to information fusion. Inf. Fusion 2006,7(1): 4-18.

[34] Stumme G, Maedche A. FCA-merge: bottom-up merging of knowledge bases. Seettle, WA, U. S., ProcIJCAI, 2001: 110-114.

[35] Abir Qasem, Dimitre A. Dimitrov etc. Efficient selection and integration of data sources for answering semantic web queries.Published in Proceedings of the Second IEEE International Conference on Semantic Computing. IEEE Computer, Society Press, 2008: 244-249.

[36] Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.Logical foundations of peer-to-peer data integration. In: Proc. of the 23rd ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems, PODS 2004:241–251.

[37] Chang C L, RC-T. Lee. Symbolic Logic and Mechanical Theorem Proving. Acad- emic Press, 1973.

[38] WANG Feiyue. Parallel Control and Management for Intelligent Transportation System: Concepts, Architectures, and Applications. IEEE Transactions on Intelligent Transportation System, 2010,11(3):630-638.