Journal of Network Communications and Emerging Technologies (JNCET) www.jncet.org
Volume 5, Special Issue 2, December (2015)

Recent Trends in Effective Design of Search Engines

Ashlesha Gupta
Assistant Professor

Ashutosh Dixit
Associate Professor

A.K.Sharma
Professor & Dean PG & Research

Abstract –Internet search engines have become an important part of everyday living and business today. Although the capabilities of Internet search engines are incrementally improving steadily, but still there are few new directions that can take the search engines to the next level. This paper summarizes the current problems in the effective search of information via WWW and explores new research trends that may be used for improving the quality of the search results by better matching the needs, preferences and intents of the user.

Index Terms – Search Engine, Crawler, Page Ranking, Semantic Web, user-Behaviour

This paper is presented at International Conference on Recent Trends in Computer and information Technology Research on 25th& 26th September (2015) conducted by B. S. Anangpuria Instituteof Technology & Management, Village-Alampur, Ballabgarh-Sohna Road,Faridabad.

1. INTRODUCTION

This WWW is a vast resource of hyperlinked and heterogeneous information including text, audio, video, image etc. that continues to grow rapidly at million pages per day. With rapid increase in information resources available via WWW and exponential growth of Internet users, it is becoming difficult to manage and access the desired information on the web. Therefore the users looking for information from World Wide Web (WWW) use search engine's interface, where they provide search queries and the results thereof are displayed instantaneously on the screen in a ranked order. A Search Engine can be viewed as a software program that takes input from the user, searches its database and returns a set of results.

In general Search Engine may return several hundreds or thousands of URL that match the keywords for a given query. But often users look at top ten results that can be seen without scrolling. Users seldom look at results coming after first or second search result page, which means that results which are not among top ten are nearly invisible for general user. When the users do not get the desired information, they modify the search query again and again till they get the desired information or get tired. The situation becomes more cumbersome when the results produced by the search tools are outdated especially when a bad URL is reported.

Owing to the problems faced by users, a need arises to develop new mechanisms that can supplement in better understanding of the user behaviour and user requirements while using search engine. This paper focuses on issues of user search trends and the problems faced in retrieving up-to-date information and try to provide an insight to the solutions that may be incorporated to improve search experience.

The rest of the paper is organized as follows: Section II gives a brief introduction of Search Engine and its working. Section III briefly explains the problems faced by users in searching information using search engines. Section IV throws some light on the reasons of deviation from providing accurate results. Section V provides an insight into the future techniques and mechanisms that may be used to improve temporal quality of search results so that the search results are more relevant and user-oriented. Section VI concludes the paper.

2. SEARCH ENGINE & ITS WORKING

A Search Engine is an information retrieval system which helps users find information on WWW by making the web pages related to their query available. Web search engines work by storing information about many web pages, which they retrieve from the WWW itself. These pages are retrieved by a Web crawler which follows every link it sees. The major components of a search engine are Crawler, Indexer and Query Processor. The typical architecture of a Search Engine is shown in Fig 1. A Crawler follows hyperlinks present in the

documents to download and store web pages for the search engine.

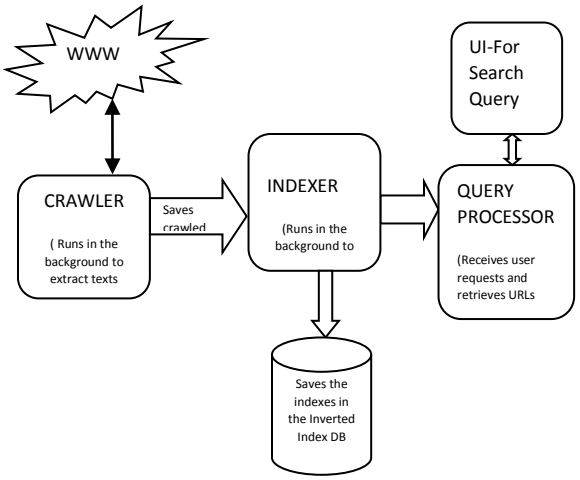

Figure 1 Architecture of Search Engine

To cover the Web as much as possible, nowadays search engines do not depend on a single but on multiple crawlers that execute in parallel to achieve the target. The contents of each page are then analyzed to determine how it should be indexed. Search Engine Indexer parses and stores data in termID, docID pairs. It extracts all the uncommon words from each page and records the URL where each word has occurred. The result is stored in a large table containing URLs, pointing to pages in the repository where a given word occurs. The purpose of storing an index is to optimize speed and performance in finding relevant documents for a search query. The query processor provides an interface for receiving and providing search requests and results to and from the user. A search engine may present a list of thousands of web pages in response to user's query possibly consisting of irrelevant web pages also. Therefore to provide better search result, page ranking mechanisms are used by most search engines for putting the important pages on top leaving the less important pages in the bottom of result list. A Page Ranker assigns relevancy scores to web pages to determine how closely a web page matches the given query. In general page ranker uses two different kinds of ranking factors: Query dependent and Query Independent for page ranking. Query-dependent are all ranking factors that are specific to a given query for example how many query words appear in the web page, how often they appear in the web page and whether the query words appear in Title or body of web page etc., while query-independent factors are attached to the documents, regardless of a given query. Some of the query dependent factors are: word document frequency, position of query terms within the document, inverted document frequency, metatags, anchor tags, language of document etc. Query independent factors considered are link popularity, click popularity, number of incoming links, directory hierarchy etc.

But in spite of having strong and efficient page ranking, indexing and search algorithms, no search engine has been able to provide accurate and precise result with reference to the search query. The users still have to scan through the returned result set to find desired information by refining it manually.

3. ISSUES IN EFFECTIVE SEARCH OF INFORMATION

The problems encountered by the user in searching information using search engines are discussed below:

- Top few pages rarely contain needed information. Relevant information pages are are extracted by scanning through multiple search pages manually.
- Most of the users feel that there are too many search results to browse, returned by the search tool i.e. the problem of information overkill.
- Sometimes users get out dated information i.e. search engine's database carries older version of web documents and not the updated one.
- Some time users get broken links in their results i.e. either the links do not exist or get modified at host server.
- Some web pages may get higher ranking because of duplicate links and self links that are meant only for increasing the popularity of the web page, but actually they do not contain any relevant information.

4. REASONS FOR DEVIATION FROM RELEVANT RESULTS

In The problems faced by the users in getting the relevant information on www are due to User Search behaviour pattern and Architecture of the current search tools. The search trends of the user that make the searching difficult are :

- Most of the users still do not use advance search features to refine the results.
- Users are still very much dependant on a single Search engine. They do not change the search tools even if they fail to get the relevant result in multiple searching.

- Most of the users prefer to modify the initial query, if they are not getting the relevant information. However, they are not aware about the related keywords of the area in which they are seeking information, leading to irrelevant results.

Constraints owing to the architecture of current search engine tools are

- Crawlers are not able to refresh the web documents on time.
- Owing to sheer size and the dynamic nature of WWW, it is almost impossible to know the exact change frequency of all web documents.
- Also the crawlers working in parallel suffer from overlapping problem in the sense that multiple copies of the same web documents may be downloaded multiple times, leading to wastage of crawler's time, network bandwidth and other resources such as storage at the Search engine side.

In the light of above discussion, it may be noted that there is a need to modify the existing techniques and/or architecture and develop new mechanisms for crawling and ranking so as to improve the quality of downloaded web documents and the result-set itself so that maximized user satisfaction may be achieved.

5. NEW RESEARCH TRENDS

5.1 Development of Semantic Web

Semantic search aims to extend and improve traditional search processes based on IR technology. These intelligent search engines incorporates Web semantics and uses more advanced search techniques based on concepts such as machine learning. The Semantic Web promises to make web content machine understandable, allowing agents and applications to access a variety of heterogeneous resources, processing and integrating the content, and producing added value output for users. . The problem of Information Overload can be partly tackled by adding intelligence to the web.

Software agents could manifest various levels of intelligent behavior from simply reactive to adaptive and learning behavior, where agents actually learn what users like and dislike. This would shield users from irrelevant information.

5.2 Opinion Mining

Improving the accuracy of search result is one of the most important objectives for Search Engines and search providers. To judge user behaviour, the user's intention needs to be identified and studied. Humans are more consistent at giving relative relevance statements. Search Engines continuously should make efforts to understand users' needs. There is a need to deploy feedback systems to capture user opinions. There is a strong relation between the relevance of Search Engine result and the satisfaction of its users for every search session. By examining the relative rating for satisfaction, it is possible to get a true user opinion about various aspects of the search session.

5.3 Focus on User Behaviour

Incorporating user behaviour data can significantly improve ordering of top results in real web search setting. Implicit feedback based on user actions can improve the accuracy of a web search ranking algorithms by as much as 31% relative to the original performance.

5.4 Development of Web Agents

Intelligent Web Agents (WA) are software programs that primarily serve two important roles: a). autonomous entities for exploring and exploiting Web-based services, and b). prototype entities for exhibiting and explaining Web-generated regularities. Web based intelligent agents are aimed at improving a Web site or providing help to a user. Personalized Multimodal Interface WA can provide users with a user- friendly style of presentation that personalizes both the interaction with users and the content presentation. This activity involves the creation of various cognitive aids, including tables, charts, executive summaries, indices, and personalized visual assistants (e.g., graphically animated personas and virtual-reality avatars). Information Gateway WA can provide users with immediate access to the most relevant information. This support encompasses a wide spectrum of information filtering and delivery activities by manipulating various heterogeneous Web sources including databases, data warehouses, newswire, financial reports, newsletters, newsgroups, outbound emails, electronic bulletin boards, and hypermedia documents, and based on users' profiles, tailoring and delivering the retrieved information to the users.

5.5 Improving web results presentation

There is need to develop techniques to radically improve the presentation of the search results by organizing them into sub-themes and including multimedia data. This has become possible because of the wide adoption of broadband in the homes, and the increasing computing power and memory capacity in the desktops and laptops. Grouping the search results in terms of various sub-themes (or facets) would often be helpful to the users, since the facets can correspond more

closely to the needs, interests, and preferences of the users. The users may select the right facets from the search results, rather than having to wade through pages of a heterogeneous mixture of the URLs and snippets of relevant results, irrelevant results, and even spam.

Table I summarizes the future techniques & their research challenges that need to be developed for improving search engine results.

Technique	Features	Advantages	Research Challenge
Semantic Data & Semantic Web	1. large interlinked database 2. Web of Ontology's , data and documents 3. uses concept of machine learning 4.technologies used are RDF, XML, OWL	1. Enable intelligent web services 2. semantically empowered search engines 3. reduced programming effort	1. construction of innovative and knowledge-based web services 2. methodological and technological support for most of the activities of the ontology development process 3. Scalability of Semantic web content.
Opinion Mining	1. Capturing public opinion about social events, political movements, company strategies,marketing campaigns, and product preferences 2. Focus on polarity detection and emotion recognition	1. Improved User Satisfaction. 2. Accurate Results. 3. opportunity for NLP researchers to make tangible progress on all fronts of NLP.	1. Create and automatically maintain and review opinion-aggregation websites. 2. designing automatic tools that crawl online reviews and condense the information gathered. 3. Future opinion-mining systems need broader and deeper common and common sense knowledge bases. 4. Deep understanding of the explicit and implicit, regular and irregular, and syntactical and semantic language rules.
Improved Result-set Presentation	1. Organizes search results into sub-themes 2. present search results in number of categories		

Focus on user feedback for Page Ranking	1. Inferred from user behavior. 2. Click-through data, Browsing features and Query-text features are considered. 2. Helps to anticipate needs and wants of clients	1. Improvement in ordering of results 2. Improved Search Engine Performance 3. Cost effective	1.Automatically predicting query difficulty, and attempt to incorporate implicit feedback for the "difficult" queries 2. exploring methods for measuring accurateness of user activities. 3. Development of user-feedback models
Web Agents	1. computational entities that are capable of making decisions 2. self improving performance 3. support for information filtering and delivering	1. Enhances search engine performance 2. reduced information overhead. 3. Personalized Multimodal Interface 4. support for collaborative computing	1. Development of multi-agent systems to attack more realistic and large scale problems 2. Developing new agent communication languages 3. Schemes to effectively allocate resources to multiple agents

6. CONCLUSION

In this article, problems of today's search engine in delivering quality result to the users are summarized. An outline of few research and development directions are given to meet the long-term challenges facing the search engines, hoping that these will lead to the next step up in the evolution and consequently, much better services to the user.

REFERENCES

[1] Bermers-Lee, T.Hall. J.A. Hendier K.O'Hara,K,"A framework for Web Science. Foundations & Trends in Web Science", 1-130, Hanover, MA, Now Publishers Inc,2006
[2] D.Lewandowski & N. Hochshtter,"Web Searching : A quality measurnment perspective ",pp(309-340), Berlin: Springer,2008
[3] B.Pan , L. Lee,"Opinion Mining & Sentiment Analysis: Foundation & Trends in Information Retrieval",2(1-2), 1-135,2010.
[4] F. Radlinski & S.Dumans. , "Improving Personalized web Search using result diversification", Proceedings of the 29th annual international ACM SIGIR conference on research and development in Information Retrieval, SIGIR'06, ACM, NY(pp691-692),2006
[5] Scott B.Huffman, "How Well Does Result Relevance Predict Session Satisfaction?". SIGIR proceedings, 2007Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.
[6] D. Kelly and J. Teevan, Implicit feedback for inferring user preference: A bibliography. In SIGIR Forum, 2003.
[7] K Jarvelin and J. Kekalainen. IR evaluation methods for retrieving highly relevant documents. In Proceedings of the ACM Conference on Research and Development on Information Retrieval (SIGIR), 2000
[8] Baeza-Yates and B. Ribeiro-Neto,"Modern Information Retrieval",. Addison-Wesley Publishing,Company, USA, 2nd edition, 2011.
[9] B. B. Cambazoglu and R. Baeza-Yates. Scalability and efficiency challenges in commercial web search engines", In Proceedings of the 36th International ACM SIGIR Conference of Research and Development in Informational Retrieval , pg 1124, 2013
[10] C. Olston and M. Najork. Web crawling "Foundationsand Trends in Information Retrieval " ,4(3):175–246,2010
[11] Ashlesha Gupta, Ashutosh Dixit, " Issues and Challenges in Effective Design of Search Engine", International Journal of Multidisciplinary Research Studies, Dec 2012.

Authors

Ashlesha Gupta is pursuing her PhD from YMCA university of Science& Technology. She did her M.E(C.E) and B.E(I.T) in the years 2008 and 2004 respectively. She is working as Assistant Professor in the department of computer engineering at YMCAUST, Faridabad Haryana. She has published various research papers in various International journals and conferences. Her research interests include Internet Technologies and Mobile and Wireless networks.

Ashutosh Dixit did his Bachelor of Engineering in CSE from RGPV MP in the year of 2001. He received his PhD and M. Tech. in Computer Engineering from MD University Rohtak, in the years 2010 and 2004 respectively. He is presently serving as Associate Professor in the department of computer engineering at YMCA University of Science & Technology, Faridabad Haryana. He has published around 80 research papers in various International journals and conferences. Presently he is supervising PhD to 6 students. His research interests include Internet Technologies, Data Structures and Mobile and Wireless communications.

Dr. A. K. Sharma, is Professor & Dean (PG & Research) BSAITM, Faridabad He completed his M. Tech. in Computer Sc. & Tech. from UOR, Roorkeeand Ph.D. in fuzzy expert systems from JMI and another Ph.D. in Information Technology from IIITM, Gwalior. Till date he has supervised more than 25 Ph.D. candidates and 100 + M.Tech. thesis. He has published 300+ papers in reputed National and International journals and conferences. He has authored 8 books.