International Journal of Engineering Research & Management Technology ISSN: 2348-4039

November- 2014 Volume 1, Issue-6

Email: editor@ijermt.org www.ijermt.org

Survey Paper on Requirement and Related Area for Improve Effectiveness of Search Engine

Manisha Mangalbhai Gujjar
M.Tech. - Student C.S.E.
Galgotias University
Uttar Pradesh

ABSTRACT:
There are major source available for information which access by use of internet. Internet has large data size access in less time requirement in single click anywhere any time. We have abundant data and information available on World Wide Web to access it we need effective search engine which provide user satisfaction for their need and application. Major problem is knowledge extraction from abundant information. There are many search engines available now days and many work related to make it more effective, but still some problem are facing by search engine to overcome it should aware of all related knowledge of search engine characteristics and as well as current trend to improvement of its performance. In this paper provide brief survey related search engines, because it play major role in knowledge century.

KEYWORD: Effectiveness of search engine.

INTRODUCTION:

Search engine are programs that search documents for specified keywords and returns a list of the documents where the keywords were found. Typical web search engine has Spider to fetch as many documents as possible then index that data for faster access and restive them as per user query satisfaction. We need search engine optimization technology to get more visibility for their sites in search engine results pages due to lots of search engine available for specific purposes. High ranking worldwide shared by general purpose Google search engine which share 80.82% market as per 2011 census., but still it facing some problem and need to overcome it in future to make it more reliable.

For search better search less. Better search result needs know about searcher, platform they using, goal of their search, content and metadata use in searching, real time updating searching, and format of data like text, audio, video, images. They need Search for particular like format, rating, source, date, location.by using metadata which is data about data provide faceted way user requirement. And manage content like web page, book, document, and object.

To make search engine more effective we should study its related anatomy like [i] user including their goals and behavior e.g. design patterns for different goals like for ask, learn, find, share, act, filter, and browse.[ii] interface to query and result including languages and affordance [iii] content including indexing, structure, metadata [iv] creator of tools, process [v]Engine performance related to features, technology and algorithms design e.g. Search algorithms are algorithms for finding an item with specified properties among a collection of items. The items may be stored individually as record in a database or may be element of a search space defined by a procedure. The combinatorial optimization is typically used when the goal is to find a sub structured with a maximum or minimum value of some parameter

Search engine should focus on different parameters like, [1] System architecture: hardware and software component, crawler, indexes, data models, and query parser e.g. Web browser is a software application for retrieving, presenting and traversing information resources on the World Wide Web. Information resources are identified by a uniform resources identifier URL. The primary purpose of a web browser is to bring information resources to the user (retrieve/fetch), allowing them to view the information (display/ rendering) and then

access other information (navigation links).web browser are built of user interface, layout engine, rendering engine, java script interface, UI backed, networking component and data persistence component.[2] Performance: how many simultaneous queries supported size of data repository and maximum sources used? E.g. Behavior of search: incremental construction, progressive disclosure, immediate response, alternate views, predictability, and recognition over recall, minimal disruption, direct manipulation. [3] File formats: types of content and data like HTML, PDF, and My SQL. Handle both structured and unstructured data e.g. Web portal is most offending one specially designed web page at a website which brings information together from diver's source in a uniform way. Types of portal are: personnel, government, cultural, corporate, hosted and many. Only authenticated and authorized user can generate requests to the application server. [4]integration: connectors, web services API for embedding search functionality in other sites and software.[5] access control: privacy, security, multiple and individual user access.[6]features: full text and metadata handle, spellcheck, wildcards, Boolean operator, ranking algorithms, query refinement, and result be stored, printed and shared.[7]implementation: installation, configuration, maintenance, vendor handle for training and support[8] pricing model: cost of ownership, service fees, by date and activity volume and unique applications and features.[9]vendor credential: positioned in market and business, financial and customer reference can we see.[10]speed: sub second response, fast answer.[11] relevance tuning: possibility to adjust ranking rates, popularity, content type, date and diversity[12] navigation and filtering: faceted search is fast for sort and limited options.[13] federated search: simultaneous search and merge of multiple database or indexes can impact on speed and performance.[14] linguistic toolset: auto categorization, entity extraction, cross walking between vocabularies and thesaurus integration.[15] search analytics: tools for measuring and understanding user behavior and find API for sharing and repositioning of this data e.g.User behavior change as per they need all relevant result or only particular results. User enjoying search or confuse and demotivate by search depend on lack of knowledge for searching or proper query generation.

Many search features control by [i]user: GUI for control interaction [ii] system: query expansion, modification, visualization, and popular query [iii] result display: ranking frequency. Give approximate or precise match for long lists for near hits. Provide probable intent of search e.g. many design patterns available for search like autocomplete, best first, federated, faceted, pagination, advanced search, personalization, structured results, actionable results and unified search.

Quality maintain: useful by solution, usable by maximum efficiency and minimum error, desirable by identify your brand , valuable by advanced the mission in competiton , accesible by all user, platform and browser , findable by site content, credible by trust, authority,popularity for most relevnt result.

Search engines vary as per general purpose, specialized domain, real time dynamic information, static page. User has to choose best from this as per their desire fulfill in less time requirement and best quality of data gain. Choosing best URLs, government and education web portals and authenticate websites give us best results as per quality and maintain privacy to safe data access.

PROBLEM DEFINAON:

Major issues are: low precision and high recall, identify user's intentions, accessible global users, and inaccurate queries.

Major challenges: generic overview of topic, invisible deep web which is not finding by spider and gives low rating. Business model require openness of internet and social collaboration. Global view information representation and Focused on providing answer rather than just a list of hits of relevant and scalable variable.

Limitations: content lacks a proper structure. Poor Interconnection of information and Automatic information transfer is lacking. Maintain trust as per many users and content access. Machine not understands the information due to lack of universal format.

Major problems related to search engines are: pay per click for advertising and popularity ranking marketing. Spam messages generation. Government uses search engines for public awareness. Editorial search require for deployment news, censor and bias. Restrict user to access specific information by some legal issues.net neutrality because ISP could charge for their search services. Irregular query expression gives unrelated information. Duplication result should filter out. Query formulation in lowest common denominator due to easily understand by machine. Slow update of internal indexes by the need of re-examine large part of the web

as well as irrelevant for transient information. We don't index every page .Privacy and data protection issues by state. Copy right issues like map maker. Search within search and protection from treats. Energy consumption for large size data manage. Specialized domain search engine exclude certain kinds of data of interest to the user. Difficult to writing search code that handle unpredictable change in the pages and data. Handle additions of new source and disappearance of other.

For development strategy including: [i] design: develop a query as for eliminate the second manual search and re-examine the result page. [ii] Implementation: co-ordinate search results when they are passing through GUI. [iii] Maintenance: continued monitoring require for dynamic nature of web, for query and result format may change and discovery of new source of information.

Proposed solution for all these problems are: the solution to click fraud and spam increasingly invasive tracking of individual user. Take hybrid approach for new generation search engine. HAKIA: provide approach for new generation search engine. Multi-agent can preprocessing of user's query before a search engine processes it. Design framework gets text input from the user in the form of complete questions, understands the input and generates the meaning. User provides the input text in English language according to his requirement after the lexical analysis of the text syntax analysis is performed on word level to recognize the world's category [i] processing natural language query [ii] information agent.

Future scope in search engine will be multisensory search: queries and results can be rendered by touch, taste, sight, sound and smell. Our matrix of hardware and software even enable the direct or indirect exchange of raw emotions. Parse data with sophisticated calculations based on own click storm. Optimize delivery of content through making the discovery process easy and ensure continued engagement. Cultural diversity of the different research disciplines emerges so need much deeper into literature of subject.

RELATED WORK:

Web based search engine: containing a high percentage of URLs pointing to the request information and a low percentage of links to poorly related or unrelated data. Development of a framework for user centered evaluation of search engine. Aggregating results from different information sources on a single result page; hence making information gathering process easier for broader topics searching. We focus on search engine whose results presenting is enriched with additional information and does not merely presented the usual list of blue 10 links. Freshness in real time and minimal relevant results set to use. Temporal document collections e.g. web archives, news, blogs, email, enterprise document. Also we look the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want. Produce more useful matches and less bad hits. Different search engine related work like; search html documents and quickly identify other pages that show the same characteristics. Grow an authorities list and almost any topic for higher quality results.

Knowledge access: formatting or indexing metadata operating system. Different visualization contexts according to the type of knowledge presented. The new knowledge discovered will in turn be added to the repositioning providing a better starting point for future data mining efforts, we need to understand if and how user information needs and search patterns very for each device. Data values for the same attributes in each record are placed into the same column in the table. Guide and allow the user to view the search results with difference perspective. Meta search engine supports search based combination keywords and search by host. Grow an authorities list and almost any topic for higher quality results. Analyze with the structure and unstructured contents of web documents. Reuse knowledge affectively without omitting useful knowledge. Simple machine learning components find the experts and aggregate their list to produce a single complete and meaningful list.

Ontology: ontology consists of vocabulary and a set of constraints on the way terms can be combined to model a domain. There are four characteristics: explicit, formulization, sharing and conceptualization. Five elements contains: class, relationship function, axiom and instance. Design patterns shall support the reuse of software architecture in different application domains as well as the flexible use of component. Components are reusable, extendable and flexible and searching across multiple distributed collections available.

Semantic web search: analyze search input query for make our search query more efficient and effective. Meaning of data on the web can be discovered not just by people but also by computers. Automated approaches

to exploiting web resources. Agent are used to perform some action or acting on behalf of a user of a computer system. Automatic data extraction technology and similarity of tag value together to extract from queries result page in structured format. Combine results in to a simple presentation require satisfying a number of layout constraints, Refinement of the multidimensional framework; use as a methodology for the evaluating of search engine from a user perspective. Agent, that helps user to find interesting content among few million.

CONCLUSION:

Search is necessity of the word as it is very difficult to get the relevant information from the information ocean. Knowledge rich data mining is the rule rather than exception. All the things presented in this paper, are necessary but also insufficient, because there so much ground to cover it. It's easy to lose sight of goal. Designer roles to repeatedly refocus attention on the user experience. In this paper, we make a brief survey of the existing literature regarding searching. We review their characteristics respectively. In the future, our work will focus on deeper and broader research in the field of search engine, with the purpose of concluding, current situation of the field and promote further development of search engine technologies.

REFERENCES:

1. Dirk Lewandowski (Hamburg University of Applied Sciences, Germany) "A Framework for Evaluating the Retrieval Effectiveness of Search Engines", available at http://www.igi-global.com/book/next-generation-search-engines/59723, pages 1-19.
2. rosalbagiugno "searching algorithms and data structures for combinatorial, temporal and probabilistic databases" submitted in partial fulfillment of the requirements for the degree of "dottore di ricerca" at universit`adeglistudi di Catania, catania, italydecember 10, 2002.
3. petermorville and Jeffery callender "search patterns" (oreilly), available at www.wowebook.com
4. Moonseo Park, Kyung-won Lee, Hyun-soo Lee, Pan Jiayi, and Jungho Yu, "Ontology-based Construction Knowledge Retrieval System" KSCE Journal of Civil Engineering (2013) 17(7):1654-1663, available at www.springer.com/12205, pages 1654-1663.
5. Jaime Arguello "Federated Search in HeterogeneousEnvironments" School of Information and Library Science, ACM SIGIR Forum, Vol. 46 No. 1 June 2012University of North Carolina, Chapel Hill, NC 27599 USA. Available at jarguello@unc.edu, 2011, pages 78-79.
6. ShanuSushmita, School of Computing Science,University of Glasgow, G12 8QQ, Scotland. "Study of Result Presentation and Interaction forAggregated Search" available at shanusushmita@gmail.com.pages 86-87.
7. NattiyaKanhabua, Norwegian University of Science and Technology "Time-aware Approaches to Information Retrieval", available at nattiya@idi.ntnu.no, page 85.
8. AndriMirzal, Faculty of Computer Science and Information Systems, University of Technology, Malaysia. "Design and Implementation of a Simple Web Search Engine", International Journal of Multimedia and Ubiquitous Engineering Vol. 7, No. 1, January, 2012, Available at andrimirzal@utm.my, pages 53-60.
9. SaniDanjuma, Jun Zhou, Hongyan Mei, AhamdAliyu, UsmanWaziri, Department of Computer Science and Technology, Liaoning University of Technology, Jinzhou, Jinzhou, China. "Design, Analysis and Implementation of Semantic WebApplications" IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 1, November 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784, Available at www.IJCSI.org, pages 110-116.
10. Matthijs van Leeuwen · Arno Knobbe "Diverse subgroup set discovery" Data Min Knowl Disc (2012) 25:208–242, DOI 10.1007/s10618-012-0273-y, available at Springerlink.com, pages 209 -242.
11. Luca Invernizzi, UC Santa Barbara, invernizzi@cs.ucsb.edu, Stefano Benvenuti, University of Genova, ste.benve86@gmail.com, Marco Cova, Lastline, Inc. and University of Birmingham,m.cova@cs.bham.ac.uk, Paolo MilaniComparetti, Lastline, Inc. and Vienna Univ. of Technology, pmilani@seclab.tuwien.ac.at, Christopher Kruegel, UC Santa Barbara, chris@cs.ucsb.edu, Giovanni Vigna, UC Santa Barbara, vigna@cs.ucsb.edu "EVILSEED: A Guided Approach to Finding MaliciousWeb Pages" 2012 IEEE Symposium on Security and Privacy, 2012, Luca Invernizzi. Under license to IEEE. DOI 10.1109/SP.2012.33, pages 428-442.
12. HaiZhuge&YorickWilks "Faceted search, social networking and interactive semantics" SpringerScience+Business Media New York 2013, World Wide Web, DOI 10.1007/s11280-013-0216-6
13. Benjamin Letham, Cynthia Rudin, Katherine A. Heller, "Growing a list" Data Min Knowl Disc (2013) 27:372–395, DOI 10.1007/s10618-013-0329-7, pages 372-395.
14. Mark Sanderson and W. Bruce Croft, "The History of Information Retrieval Research" Proceedings of the IEEE | Vol. 100, May 13th, 2012
15. Hsiang-Yuan Hsueh, Chun-Nan Chen and Kun-Fu Huang, "Generating metadata from web documents: asystematic approach" Hsueh et al. Human-centric Computing and Information Sciences 2013, 3:7, available at http://www.hcis-journal.com/content/3/1/7, pages 1 to 17.

16. Raghu Ramakrishnan, Bee Chung Chen, "Exploratory mining in cube space", Springer Science+Business Media, LLC 2007, Data Min Knowl Disc (2007) 15:29–54 DOI 10.1007/s10618-007-0063-0, pages 29-54
17. Ero Balsa, Carmela Troncoso and Claudia Diaz ESAT/COSIC, IBBT KU Leuven, Leuven, Belgium "OB-PWS: Obfuscation-Based PrivateWeb Search" available at firstname.secondname@esat.kuleuven.be, pages 491 to 505.
18. Neil Y. Yen, Timothy K. Shih, Senior Member, IEEE, Louis R. Chao, and Qun Jin, Member, IEEE "Ranking Metrics and Search Guidancefor Learning Object Repository" IEEE TRANSACTIONS ON LEARNING TECHNOLOGIES, VOL. 3, NO. 3, JULY-SEPTEMBER 2010, pages 250 to 264.
19. Muhammad Sajid Khan, Chengliang Wang , Ayesha Kulsoom, ZabeehUllah, "Searching Encrypted Data on Cloud", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 1, November 2013, ISSN (Print): 1694-0814 | ISSN (Online): 1694-0784, available at www.IJCSI.org, pages 230 to 233.
20. Greg Goth, "A Structure for Unstructured Data Search" January 2007 (vol. 8, no. 1), art.no. 0701-o10031541-4922 © 2007 IEEE, Published by the IEEE Computer Society
21. Pedro Domingos, "Toward knowledge-rich data mining", Data Min Knowl Disc (2007) 15:21–28 DOI 10.1007/s10618-007-0069-7, pages 21 to 28.
22. MaristellaAgosti, Franco Crivellari, Giorgio Maria Di Nunzio, Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction, Data Min Knowl Disc (2012) 24:663–696, DOI 10.1007/s10618-011-0228-8, pages 663 to 696.
23. Myra Spiliopoulou, BamshadMobasher, OlfaNasraoui, OsmarZaiane, "Guest editorial: special issue on a decade of mining the Web" springer, pages 273 to 277.
24. Shridevi Swami1, PujashreeVidap, "Web Scraping Framework based on Combining Tag and Value Similarity", Department of Computer Engineering, Pune Institute of Computer Technology, University of Pune, Maharashtra, India, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013, available at www.IJCSI.org, pages 118 to 122.
25. jotsna Molly Rajan, M. Deepa Lakshmi, "Ontology-based Semantic Search Engine for Healthcare Services" International Journal on Computer Science and Engineering (IJCSE), jotsna@gmx.com, deepasuresh12@gmail.com, pages 589-595.
26. S.G.Choudhary, S.R.Kalmegh, Dr. S. N. Deshmukh, "Semantic Search Algorithms based on Page Rank and Ontology: A Review", 3rd International Conference on Intelligent Computational Systems (ICICS'2013) January 26-27, 2013 Hong Kong (China), pages 17-20.
27. Wenjuan Wang, Huajuan Mao, Weihui Dai, Yiming Sun, "Technological Resource Search Engine Based on Ontology", Proceedings of the Third International Symposium on Electronic Commerce and Security Workshops(ISECS '10), Guangzhou, P. R. China, 29-31,July 2010, pp. 124-127.
28. AnkitKanojia and Varsha Sharma, "University Search Engine Based on Semantic Approach", International Journal of Computer Theory and Engineering, Vol. 4, No. 4, August 2012, pages 497-500.
29. M. AsifNaeem, Noreen Asif, "A Web Smart Space Framework for Intelligent Search Engines", Department of Computer Science, University of Auckland, NewZealand., International Journal of Emerging Sciences ISSN: 2222-4254 1(1) April 2011, mnae006@aucklanduni.ac.nz, kinzajameel@yahoo.com,
30. S. A. Inamdar1 and G. N. Shinde, "An Agent Based Intelligent Search Engine System For Web Mining" 1School of Computational Sciences, Swami RamanandTeerthMarathwada University, Nanded-431606, INDIA, 2Indira Gandhi College, CIDCO, Nanded-431603, INDIA, Research, Reflections and Innovations in Integrating ICT in Education.
31. Ralph E. Johnson, "Frameworks (Components+Patterns)", How frameworks compare to other object-oriented reuse techniques, October 1997/Vol. 40, No. 10 COMMUNICATIONS OF THE ACM, pages 39-42.
32. G.Madhu and Dr.A.Govardhan, Dr.T.V.Rajinikanth, "Intelligent Semantic Web Search Engines: A Brief Survey", International journal of Web & Semantic Technology (IJWesT) Vol.2, No.1, January 2011, available at madhu_g@vnrvjiet.in, govardhan_cse@yahoo.co.in, rajinitv_03@yahoo.co.in.
33. Doh-Shin Jeon, Bruno Jullienand Mikhail Klimenko, "Language, Internet and PlatformCompetition: thecase of Search Engine", yToulouse School of Economics (GREMAQ, IDEI) and CEPR.dohshin.jeon@gmail.comzToulouse School of Economics (GREMAQ and IDEI). bruno.jullien@tse-fr.eu xGeorgia Institute of Technology. Mikhail.Klimenko@econ.gatech.edu
34. david hawking, nick craswell, "measuring search engine quality",
35. John Davies, Alistair Duke, Nick Kings, DunjaMladenic´ , KalinaBontcheva, MihaGrc˘ar,
36. Richard Benjamins, Jesus Contreras, Mercedes BlazquezCivico and Tim Glover, "Next generation knowledge access", JOURNAL OF KNOWLEDGE MANAGEMENT j VOL. 9 NO. 5 2005, pp. 64-84, Q Emerald Group Publishing Limited, ISSN 1367-3270, pges 64 to 84.
37. RanjeetDevarakonda, Les Hook, GiriPalanisamy, Jim Green, "Next-Generation Search Engines for Information Retrieval", International Journal of Software Engineering (IJSE), Volume (2) : Issue (1) : 2011, available at jgreen@iiaweb.com.
38. NEXT GENERATION SEARCH, July 2010, icsti insights, pages 1 to 30.
39. F C Johnson, J R Griffiths and R J Hartley, "DEVISE, A framework for the evaluation of Internet search engines", The Council for Museums, Archives and Libraries, 2001, cerlim.