Volume 3, Issue 4, April 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Study on Information Retrieval Proficiencies for Mining the Network

[1]Dheeraj Agarwal,[2]Prateek Sharma,[3]Jitendranath Srivastava
[1]Research Scholar Invertis University, Bareilly, India.
[2]Research Scholar Invertis University, Bareilly, India.
[3]Associate Prof Invertis University,Bareilly, India.

Abstract – Internet has issued as one of the spiritualist for acquiring information regardless of the query submitted by search engines. There exist several commercial web search engines to capitalize the demand of such information learning task. Such programs crawl the web and retrieves twisty information in much faster way. In-spite the information gathered is voluminous, the problem of relevancy is still a major issue even today and hence the retrieved information in an ineffective in nature and leads to poor performance. Lots of research work has been carried out over the years and however there is an uphill task to retrieve relevant results capitalizing the user's task. This paper furnishes a brief survey on different cases of web mining accesses namely content mining, link mining and structure mining. Each of these approaches was looked into with relevant background study in a detailed fashion and finally the paper resolves by demonstrating a comparative survey to analyze the effective approach among the existing information recovery techniques, there by presenting the results that satisfy the web surfer.

Keywords – Web mining, link mining, search engine, search query, precision, and recall

1. Introduction

Search engines have no limitations on the capability as far as language is concerned. There are few works exploring the capabilities of non-english language search like Russian, French, Hungarian and Hebrew [1]. Google is very popular in non-English speaking countries and is widely accepted by research community. Search procedures for efficient searching vary by different levels ranging from text search [21], image search, audio and video [2]. Acquiring desired selective information from such a search engine is ambitious due to varying level of expertise and content representation. There exist various approaches like personalization of search processes, which could enhance the web search procedure. Such a task involves manual interruption which may be automatized or manual based on content, structure or link mining. Lot of research work has been accomplished in several dimensions and this paper surveys on these advances and finds the best suited approach for further exemplification. In spite of the advantages that a search process provides, there are several issues which leads to irrelevant search process are i) Unknowingness of search procedure ii) Search engines not realizing the query which needs query reformist iii) Query expansion using semantic analysis using WordNet, NLP tools etc. Also there have been few attempts focused on enhancing the search results constituting the context [3]. To overcome the discrepancies furnished by search engines, like adopting agents [4], rectangular and signed approach [15], document classification [5], ontology [8] and semantics [10, 18]. Over the years, research on web mining has been extended to cover the use of data mining and similar techniques to discover patterns or that are hidden from the Web-related data such as Web usage data or Web server logs. In spite of such complexities, the freshness of a data update still makes the process more tricky and challenging [6]. The task also requires careful design of crawling process [7].

The paper is coordinated as follows. Section 1 has provided the basic function of a search engine and the focus of the paper. Section 2 presents the background study and finally section 3 provides the challenges put forth and finally section 4 provides conclusion.

2. Background study:

Louise F. Spiteri [9] built a comparative study using six different types of search engines in on-line shopping domain depending on a number of factors, such as the quality, variety and price of their products, their guarantees, return policies, etc. The success of these commercial web sites is connoted upon the more basic assumption that consumers can actually find the sites on the World Wide Web. A major advantage of online shopping is that it enables consumers to engage in comparability shopping with an ease that cannot be repeated easily in the physical world. The authors have also studied the ways in which these Internet search engines facilitate access to online shopping sites via their hierarchical subject lists. Specifically, the authors examined the internal structure, consistency and predilection of the six subject lists. The findings suggest that the search directories (i) use equivocal and sometimes misleading categories to organize e-commerce sites, (ii) are only reasonably consistent in the way they organize e-commerce sites and (iii) provide comparatively few opportunities for comparison shopping. Adnan H. Yahya and Ali Y. Salhi [11] have demonstrated the challenges that the Arabic language brings in for retrieving Web information and some Arabic NLP tools. The authors have formulated techniques for returning good search consequences by helping search engines better

© 2013, IJARCSSE All Rights Reserved

realize users' queries and adding features to what presently exists in search engines. This paper described on our work on designing text mining and query pre-processing tools that are able to efficiently process and search large quantities of Arabic web data. We applied a statistical/Corpus-based approach, and constructed databases that contain Arabic words from newspapers with their frequencies and used that as basis to create well structured search aid tools that are able to handle arabic content and which are capable of being incorporated into existing web search engines and document processing systems. Munish Kumar [12] described a large scale evaluation of the effectiveness of HITS, SALSA, sNorm (p) algorithm comparisons with other link based ranking algorithms. Author has carrie out the evaluation with respect to a large number of human evaluated queries using major measures of effectiveness: MRR and MAP [11]. The measurement presented in work provides a solid evaluation of the best well-known ranking algorithm.

Joran Beel et al [13] introduced and discussed the concept of academic search engine optimization (ASEO). Based on three recently conducted studies, the authors have provided guidelines on how to optimize scholarly literature for academic search engines in general and for Google Scholar in particular.

Parul Gupta and A.K.Sharma [14] concentrated on to provide most relevant documents to the users in minimum possible time by granting effective and fast accesses by indexing the web pages after they have been assembled into a repository by the crawler. The work proposed by the authors focus on to index that is built on the basis of context of the document rather than on the terms basis using ontology. The ontology-based accumulation selection method presented seems to use the context to describe collections and search engines. The context of the documents being gathered by the crawler in the repository is being extracted by the indexer using the context repository, thesaurus and ontology repository and then documents are indexed according to their respective context.

Sumalatha Ramachandran, Sharon Joseph, Sujaya Paulraj and Vetriselvi Ramaraj [16] suggested an optimal Load Shedding algorithm which is used to handle overburden conditions in real-time data stream applications and is adapted to the Information Retrieval System of a web search engine thereby providing a trustworthy search results to the user within an optimum response time, even during overburden conditions.

Semantic technologies predict a next generation of semantic search engines, taking into consideration the semantic relationships between query terms and other concepts that might be significant to user. A. M. Riad et al [17] have presented a general framework for personalized Semantic Search Engine (PSSE). PSSE is a crawler-based search engine in which multi-crawlers work in parallel to traverse both traditional as well as semantic web. Additionally, user interests and preference are automatically learned from Web usage data and integrated with page authoritativeness and content relevancy to rank final results.

Adish Singla et al [19] quantified the benefit that users currently obtain from trail following and comparing different methods for finding the best trail for a given query and each top-ranked result. Then the authors compare the relevance, topic coverage, topic diversity, and utility of trails selected using different methods, and break out findings by factors such as query type and origin relevance. The findings demonstrate value in trails, highlight interesting differences in the performance of trail finding algorithms, and show that the found best-trails for a query outperforms the trails most users follow. Also the findings have implications for enhancing Web information seeking using trails.

Kavita D. Satokar and Prof.S.Z.Gawali [20] presented a personalize Web search system, which can helps users to get the relevant web pages based on their selection from the domain list. Users thus obtain a set of interested domains and the web pages from the system. The system is based on features extracted from hyperlinks, such as anchor terms or URL tokens, user interest domains and past search results. The authors proposed method uses an innovative weighted URL Rank algorithm based on user interested domains and user query.

Kamlesh Patidar et al [22] presented an approach for multi-level data tracking and logging. The authors have introduced a new approach, called multiple layered database (MLDB) approach using webmart and investigated for resource discovery and content accessibility. The approach is to construct content digital library in which accessibility of content is as simple as accessing content of word from index of a book. These data is cleaned from unnecessary information (i.e. requested images) in the preprocessing task, followed by the pattern discovery process that generates valuable information about user behavior. The major strength of the MLDB approach provides a tight integration of database and data mining with resource discovery and content accessibility from repository of objects. Additionally, multi-level data tracking methodology approach was introduced.

Lihui Chen and Wai Lian Chue [23] depicted a novel representation technique which makes use of the Web structure together with summarization proficiencies to better represent knowledge in actual Web Documents. The authors have named the suggested proficiency as Semantic Virtual Document (SVD). The authors also drafted an observational design to evaluate the effectiveness of the proposed SVD for representation and presented a prototype called iSEARCH (Intelligent SEarch and Review of Cluster Hierarchy) for Web content mining.

Sauparna Palchowdhury et al [25] looked into the information seeking process in a specific manner by including negative constraints, i.e., specifications of what the user is not looking for. A search engine that builds use of such constraints is likely to return more accurate results. The authors have considered the problem of identifying such negative constraints from verbose queries. A maximum-entropy classifiers trained to describe negative sentences in verbose queries with about 90% accuracy. In the next study how retrieval strength is affected when these negative sentences are excreted from the queries and the results were found to provide modest advances in retrieval accuracy.

Gamal F. El-Hady et al [24] have confronted a method for advocating the related queries to the input based on clustering process over the web queries extracted from a search engine query log. The proposed method is based on clump processes in which groups of semantically similar queries are discovered. The clustering process uses the content of historical orientations of users registered in the query log of the search engine. This facility furnishes queries that are

© 2013, IJARCSSE All Rights Reserved

related to the ones presented by users in order to direct them toward their required information. Also the method not only discloses the related queries but also ranks them allotting to a similarity measure.

3. Challenges & focus:

Currently, all existing search engines adopt several techniques & approaches as listed in above section with the motive of improving the performance of Web search engines. From the above literatures presented, we are aware that web mining approaches are broadly categorized as web content mining, web structure and web usage mining. Though all these categories were investigated to a modest level, mining the web using content mining achieves remarkable importance, where as interest on structure mining and usage mining has been a bit lower as compared to deep content mining. Also, as detailed in the survey, few attempts were made to restructure the query, providing alternate queries or personalizing web search and providing negation terms that could exclude the terms during search process are notable.

Diversified approaches to optimize search procedure, ranking based on importance, summarization have drafted huge attention among research community. Evaluating the performance of search engine based on retrieved web contents seems to be a daunting task. However, since the evaluation is subjective and depends on users' knowledge and interests, we may get different evaluation results from different users. To evaluate we need to evaluate the summaries based on the target set created manually by professionals. Such target generated by humans vary depending the task they have been assigned. The following are the challenges which are wide open:

 Efficient searching procedure/strategies

- Evaluation of obtained documents
- Generating the target seed
- Identifying of better search engine

With the exponential growth rate of online data and recovering the apt information demanded from millions of document repositories, research on web mining still continues to be a huge issue. Our work focus on to mining the web using content and we focus on to receive best results during search process.

4. Conclusion:

This paper has demonstrated a brief survey on various approaches/techniques for amending the search results remembered by search engines. A brief analysis on different web mining approaches was demonstrated. The challenges and the scope were presented further, so and the focus of the paper is clearly narrowed.

References

[1] Judit Bar-Ilan and Tatyana Gutman, "How do search engines respond to some non-English queries?", Journal of Information Science, 31 (1) 2005, pp. 13–28.

[2] Guojun Lu, Ben Williams and Chooneng You, "An effective World Wide Web image search engine", Journal of Information Science, 27 (1) 2001, pp. 27–37.

[3] Jesús Vegas, Fabio Crestani and Pablo de la Fuente, "Context representation for web search results", Journal of Information Science, 33 (1) 2007, pp. 77–94.

[4] Huaiqing Wang, Stephen Liao and Lejian Liao, "An agent-based framework for Web query answering", Journal of Information Science, 26 (2) 2000, pp. 101–109.

[5] Offer Drori and Nir Alon, "Using document classification for displaying search results lists", Journal of Information Science, 29 (2) 2003, pp. 97–106.

[6] Dirk Lewandowski, Henry Wahlig and Gunnar Meyer-Bautor, "The freshness of web search engine databases", Journal of Information Science, 32 (2) 2006, pp. 131–148.

[7] Mike Thelwall, "A web crawler design for data mining", Journal of Information Science, 27 (5) 2001, pp. 319–325.

[8] Ying Ding, "A review of ontologies with the Semantic Web in view", Journal of Information Science, 27 (6) 2001, pp. 377–384.

[9] Louise F. Spiteri, "Access to electronic commerce sites on the World Wide Web: an analysis of the effectiveness of six Internet search engines", Journal of Information Science, 26 (3) 2000, pp. 173–183.

[10] Ying Ding, "Semantic Web: Who is who in the field - a bibliometric analysis", Journal of Information Science, 36 (3) 2010, pp. 335–356.

[11] Adnan H. Yahya and Ali Y. Salhi, "Enhancement Tools for Arabic Web Search A Statistical Approach", In Proceedings of International Conference on Innovations in Information Technology, pp.71-76.

[12] Munish Kumar, "A New Approach for Web Page Ranking Solution: sNorm (p) Algorithm", International Journal of Computer Applications, Vol. 9, No.10, pp.20-23, November 2010.

[13] Joran Beel, Bela Gipp, and Erik Wilde, "Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co.", Journal of Scholarly Publishing, 41 (2): 176–190, January 2010.

[14] Parul Gupta and Dr. A.K.Sharma, "Context based Indexing in Search Engines using Ontology", International Journal of Computer Applications, Vol.1, No. 14, pp.49-52.

[15] G.Poonkuzhali, G.V.Uma and K.Sarukesi, "Detection And Removal Of Redundant Web Content Through Rectangular And Signed Approach", International Journal of Engineering Science and Technology, Vol. 2, Issue.9, 2010, pp. 4026-4032.

© 2013, IJARCSSE All Rights Reserved

[16] Sumalatha Ramachandran, Sharon Joseph, Sujaya Paulraj and Vetriselvi Ramaraj, "Handling Overload Conditions in High Performance Trustworthy Information Retrieval Systems", Journal of Computing, Vol. 2, Issue. 4, pp.70-75, April 2010.

[17] A. M. Riad,Hamdy K. Elminir, Mohamed Abu ElSoud, Sahar. F. Sabbeh, "PSSE: An Architecture For A Personalized Semantic Search Engine", International Journal on Advances in Information Sciences and Service Sciences, Vol,2, No. 1, pp.102-112, March 2010.

[18] S. Raja Ranganathan, Prof. M. Sadish Sendil and S. Karthik, "Relation Based Semantic Web Search Engine", International Journal of Academic Research, Vol. 2. No. 3, pp. 96-100, May 2010.

[19] Adish Singla, Ryen W. White and Jeff Huang, "Studying Trailfinding Algorithms for Enhanced Web Search", In Proceedings of ACM SIGIR'10, pp. 443- 450, July 19–23, Geneva, Switzerland.

[20] Kavita D. Satokar and Prof.S.Z.Gawali, "Web Search Result Personalization using Web Mining", International Journal of Computer Applications, Vol. 2, No.5, pp. 29-32, June 2010.

[21] G.Poonkuzhali, R.Kishore Kumar, R.Kripa Keshav, K.Thiagarajan and K.Sarukesi, "Effective Algorithms for Improving the Performance of Search Engine Results", International Journal Of Applied Mathematics and Informatics, Vol. 5, Issue 3, pp. 216- 223, 2011.

[22] Kamlesh Patidar, Preetesh Purohit and Kapil Sharma (2011), "Web Content Mining Using Database Approach and Multilevel Data Tracking Methodology for Digital Library", International Journal of Computer Science and Technology, Vol. 2, Issue. 1, pp.194-198, March 2011.

[23] Lihui Chen and Wai Lian Chue, "Using Web structure and summarisation techniques for Web content mining", Information Processing and Management, Vol. 41 , pp. 1225–1242, 2005.

[24] Gamal F. El-Hady, Hamada M. Zahera and W. F. Abd El-Wahed, " Query Recommendation for Improving Search Engine Results", International Journal of Information Retrieval Research, 1(1), 45-52, January-March 2011.

[25] Sauparna Palchowdhury, Sukomal Pal and Mandar Mitra, "Using Negative Information in Search", In Proceedings of Second International Conference on Emerging Applications of Information Technology,2011, pp.53-56.

© 2013, IJARCSSE All Rights Reserved