

“Qualitative Approaches in Content Mining - A Review”**Abdul Karim Siddiqui¹ and Ashish Sutte²***School of Computer Applications, Lovely Professional University, Punjab**School of Pharmaceutical Sciences, Lovely Professional University, Punjab***Abstract**

We are living in a digital world which is full of information. People are dependant over the information that is concerned with them. The Tour and Travel companies are modifying destination's preferences as they are paying high SEO costs to attract new customers. Hospitals are prescribing a lot of Medicines and Health Mantras and conversely Yoga and Meditation have been evolved to give you a new feelings; the feelings that guarantee eternal happiness! All these happen online because you are digital a guy. They know how to make money from you E- wallet. Your each day passes with a lot of online transactions and moving your fingers on touch screens and pressing keyboards. Today's online environment provides dynamic course materials to learners according to their needs and preferences. In case of quality content mining, it should be more précised and accurate. Intentional attempts may be made to manipulate search engine rankings by targeting and stuffing specific keywords or keywords phrase reached queries. Quality Content availability and its accessibility on top search ranking is big issue before us. Data Mining being the subject of data analysis to discover the pattern in data and relationship with other correlated data, deals and makes valid prediction. These happen by classification, clustering, prediction, association, extraction and sequence detection of various types of problems [1]. Through this paper, an endeavor is made to survey better digging methodologies for immense data mass accessible on the web.

Keywords: Data Mining, Quality Content Mining, Search Engine Optimization.

1. Introduction

Mark R. Patterson [2] in his paper tentatively suggests an evaluation by some general observations regarding finding useful information products in two-sided markets. The paper sets some Allegations on Google i.e. it has manipulated searched listings. The searched listed links are shown either in “organic” search results or the results by AdWords sponsored-advertisings. It was also assumed that Google did it to disadvantage competitors or potential competitors. The allegations are that Google does not deliver search results that may produce quality content to consumers. These practices lead to Market Power to acquire advantages in their own favour. These force to compromise the quality of content, a visitor need to have from search engine’s results. It becomes tedious for user to assess, the result and to determine which one is best suited for him. It should be further confirmed that Search engines provide search results for free. There is a need to develop an intelligent computing environment based quality content mining system that integrate soft computing and probabilistic based techniques like Artificial Intelligence, Data Mining (DM), and Machine Learning to provide adaptive learning content according to learner preferences. As a sign of best approaches, this paper tries to survey the developments made in the direction of better content mining technique by evaluating different work and progress made in this field. Most of developments have been reviewed from 2005 to 2019. This period is considerable as it covers some better reviewed ideas for reliable data mining techniques. To be focused on Quality Content Mining, with rapidly changing technologies, it has utmost a spontaneous part particularly in the light of approaches for accessing data from database servers. The input for this literature review, begun with search made through the keyword indices on the Springerlink , Science Direct, IEEE Xplore and other online databases, for contents matching with article possessing the phrases “A study on quality ” of “Qualitative Approaches in Data Mining Process.”

2. Search Engine Marketing

Danny Sullivan in 2001 proposed the term "Search Engine Marketing" to explain functionalities required in performing SEO, showing paid listings, directory submissions, developing online marketing strategies, organizations, and for individuals. Search Engine Marketing indicates set of activities including Optimization Processes. It aims for return on investment (ROI). Aggregating relevant traffic building may be secondary. Web Pages are traced by a software program called or a Internet bot. It systematically browses the WWW, for Web indexing. Exclusions of crawling can be made on one or more WebPages by the use of robots.txt. The metadata i.e. Data about these crawled web pages are kept in an index database so that further queries may be sorted out. There may be several demands of political, economic or social pressure that can affect the outcome shown by Search Engines

[3].

- Paid Search is made when a business pays amount to Search Engines in order to stand at the top of search outcomes. It is so attributed to catch the attention from consumers clicking on a link in or an advertisement and sponsored listing [4]. Google Internal Search Data shows that Paid client base will grow multiple times more than 1.6 million clients to 9.6 million clients in India by 2020.
- When anyone tries to search any online content(s) by putting any keyword(s) or queries on Search Engines. The Search traffic flashes from various business links among search engine result pages. These search listing can be influenced by using SEO online/offline Tools. These types of listing of links that appear below the ads are known as "organic results." See Fig-1 Paid and Organic Searching.

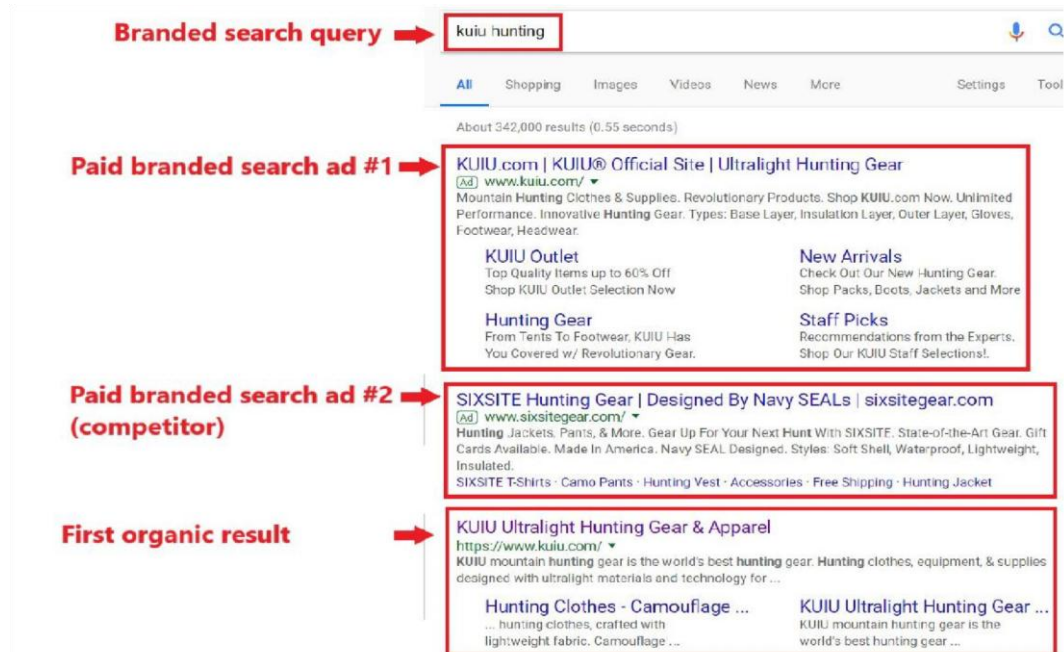


Fig. 1 – Paid and Organic Searching

3. Different Approaches in Content Mining

The potentially useful and previously unknown, unexplored information or knowledge from the ocean of data over Web(s) can be extracted from Web content mining. It may be understood as a converging research area from several research fields in Computer Science and Applications such as database, information retrieval and AI especially from machine learning and NLP. Some of approaches made in support of Web Content Mining under Data Mining best techniques are reviewed as follows:

[i] The proposed algorithm given by G. Poonkuzhali [5], explored the benefits of full word matching and signed approach by using organized domain dictionary. The word length decides the Indexing of data. The proposed theory claims for better utilization of memory space and searching time. Through the paper the author points the efficiency and quality of the search engines.

[ii] Rahmani, R. et al [6] presented a localized content-based image retrieval (CBIR) system. A user may be interested in fragmented image rather reading whole object. The authors suggested a CBIR system- ACCIO. The CBIR has a limitation in representing the image to capture the quality Inputs. The authors introduced some novel image representations technique which are the traditional segmentation-based and salient point-based technique.

[iii] Mehdi Hosseini et al [7], proposed an approach to investigate information architecture and web site link. The proposed model is termed as Mining search engine query log. To support their work, they proposed a novel query-URL co-clustering technique. For monitoring and evaluate the data sets of different web sites, it has been fetched with a massive request log.

[iv] Ricardo Campos et al [8], proposed an approach that considered both the lexical and semantics similarities in documents. To produce semantically meaningful clusters, an activation spreading technique has been applied. A clustering mechanism that forms the semantics of the retrieved documents is basically used.

[v] P. Chahal et al [9] “Ranking of Web Documents using Semantic Similarity” : They proposed a superior plan to make one's hunt information very legitimate and productive. The method gives connection between looked through outcomes and client questions. They investigated important relations between the catchphrases, recommending client's goal and to decide their significance on recorded web(s).

[vi] P. Rani et al[10], in their research work, they proposed a new algorithm to calculate web page rank called M-HITS. M-HITS can be pronounced as new version of HITS

algorithm. In this algorithm six different parameters were used to evaluate rank for web pages. This technique may prove fruitful for Future work by using some AI techniques as well.

[vii] Raymond Kosala et al [11], in “Web Mining Research: A Survey”, they analyzed the region of Web content mining and called attention to certain disarrays with respect to the utilization of the term 'Web mining'. They additionally recommended diverse Web mining classifications.

[viii] Seema Rani et al [12], Proposes a system which presents an improved Semantic Similarity technique with AI to rank the WebPages. They tested it over 1000 web pages under various categories viz. computer, education, chemistry, programming, etc. Results were checked on various inputs on the basis of queries supplied. The accuracy resulted up to 95%. The proposed System can produce better result if is implemented on the cloud servers. A multithreading technique can also be fruitful to save time. To improve results, it can be applied on categorized Web Pages.

4. Data Mining – An Overview

Across all the known fields, data resides at centre. It may be collected, accumulated and presented as more satisfactory way. Data is the fuel of this era; obviously Data Mining plays a vital role to explore right information from huge data source. This is the way of extracting genuine information from huge data source available over Internet. Today the advancement in varieties of data, forces us to develop a system of effective computation theories and related tools to guide humans in mining desired knowledge. Today we have several applications of more précised data mining techniques for pattern discovery [13]. The methods for data mining range from - a. generalization, b. characterization, c. classification, d. clustering, e. association, f. evolution, g. pattern matching, h. data visualization and i.

meta-rule guided mining. [14]. Today Computer based Information System is applied in Data Mining process as new techniques, for discovering knowledge from huge database [15].

Disciplines Involved in Data Mining

Data mining: “It is the process of examining large pre-existing databases to generate new information.” Web mining is subset of data mining. Web mining is additionally partitioned into two subsets - Web content mining and Web use mining. Content mining centers around the substance of a website page. Web utilization mining finds the character or inception of Web clients. Use of clients perusing conduct at a Web website is likewise caught by Usage Mining. Mining of aural and visual information, for example, Sound, Video Content and Images have likewise been slanting out in the open requests.

5. Web Spamming

When web pages are crafted to gain higher ranking of some or all affiliated web pages without improving the utility to viewers, the cases of Web Spamming arise. It's a big market to achieve a higher rank in search engine result pages by provision of paid inclusions. Google surveys indicate that in India 63% growth has been noticed in queries for smart phones whereas 42% growth in queries for personal computers during 2017.

From multinational companies to local brands are now aiding themselves by SEO market. Thus, a multi-billion dollar industry has been blossomed in search engine optimization (SEO). The market of SEOs has been growing rapidly. Web spam can be noticed into many forms as indicated in figure – 2. The below figure shows how incoming links can generate spamming cases during Internet use. This may also cause in various forms as outgoing links may contains hidden interests[17-19].

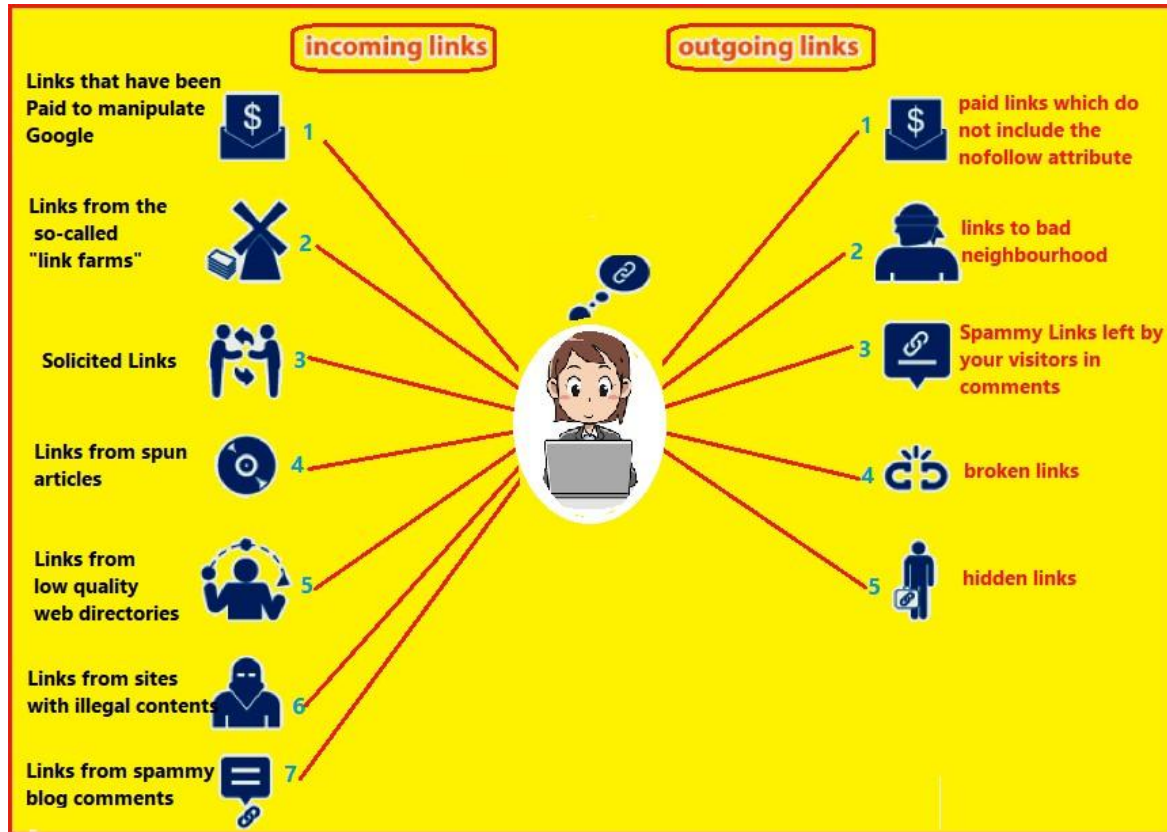


Fig. 2- Web Spamming

6. Discussion

The accuracy of data mining methods depends upon the size and features of observed data set. The data sets may be highly imbalanced; due to the prediction errors when run by the classifiers. Other problem may be missing values in data sets. The sample size may also affect the analysis. So, that there seems no proper data mining method to balance all quality mining issues. The time has now been shifting towards stuffing of SEO techniques which can dilute your searched query as per companies' agenda. There also, the chance of Web spamming has now been increasing to steal your search habits and companies are making money by analyzing the pattern and the product you search for. As per Google Webspam report 2018, 180,000 search spam reports were submitted by users (double the figure from 2017), 64 percent of which were acted upon. Google covers a big proportion in Search

engine market shares. It has initiated a Knowledge Graph concept to bind people's search expectations where a knowledge base is used to enhance Google Search Engine's search results through semantic search information, collected from a wide range of sources, so as to enable searching for things, buildings, people and landmarks that Google knows about. You get your requested information instantly. This may be proved as revolutionary step towards crafting the next generation of search which would be right synonym for quality content mining. This collective intelligence of the webs will sketch quality search.

7. Conclusion

There is still required a better effort to draw a hybrid model which could manage the issue discussed above. The users of Internet should have the right of getting things of matter (TOM). A survey from Google says that in India By 2021 Online education will become a \$2 billion industry growing at the pace of 8 times. Our future direction may lead to enhance and enrich our predictions using hybrid models. Mining quality contents is our right. Web Mining needs to be developed in such a way so that people may choose what is best amongst available resources. The time and cost also matter here. This new age of Content Mining is broadly interdisciplinary, attracting researchers of artificial intelligence, machine learning, databases, IR, cognitive social theory and behavioral science. Our thrust to search quality contents should not stop in front of manipulated search results. The systems for E-content mining and benefitting common people by availing low cost and quality search information should be in our priorities.

References

- [1] Minky Jindal & Nisha Kharb, "Data Mining in Web Search Engine Optimization and User Assisted Rank Results", International Journal of Computer Applications (0975 – 8887) Volume 95– No.8, June 2014.
- [2] Mark R. Patterson, "GOOGLE AND SEARCH-ENGINE MARKET POWER"

- [3] Dr. Tom Seymour , “History Of Search Engines” International Journal of Management & Information Systems – Fourth Quarter 2011 Volume 15, Number 4 , Minot State University, USA
- [4] Battelle, John. The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture. New York: Portfolio. (September 2005).
- [5] G. Poonkuzhali, K.Thiagarajan et al “Signed Approach for Mining Web Content Outliers”, World Academy of Science, Engineering and Technology 56 2009
- [6] Rahmani, R.; Goldman, S.A.; Hui Zhang; Cholleti, S.R.; Fritts, J.E.; , "Localized Content-Based Image Retrieval," Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.30, no.11, pp.1902-1912, Nov.2008
- [7] Mehdi Hosseini , Hassan Abol hassani , “Mining Search Engine Query Log for Evaluating Content and Structure of a Web Site” in Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence.
- [8] Ricardo Campos, Gael Dias, Celia Nunes, "WISE: Hierarchical Soft Clustering of Web Page Search Results Based on Web Content Mining Techniques," wi, pp.301304, 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), 2006.
- [9] Poonam Chahal, Manjeet Singh & Suresh Kumar, “Ranking of Web Documents using Semantic Similarity”, International Conference on Information Systems and Computer Networks, 2013, Pg 145-150.
- [10] Parveen Rani and Er. Sukhpreet Singh, “An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters” (IJCT) ISSN: 2277 – 3061, Vol 9 No 1 (2013)
- [11] Raymond Kosala and Hendrik Blockeel, “Web Mining Research: A Survey”, kuleuven.ac.be Volume 2, Issue 1 - page 15

- [12] Seema Rani, Upasana Garg, “ A Ranking Of Web Documents Using Semantic Similarity And Artificial Intelligence Based Search Engine” International Journal of Science, Engineering and Technology Research (IJSETR), Volume 3, Issue 12, December 2014
- [13] R. Agrawal and G. Psaila, “Active data mining,” Current, pp. 3–8, 1995.
- [14] S. H. Liao, P. H. Chu, and P. Y. Hsiao, “Data mining techniques and applications - A decade review from 2000 to 2011,” Expert Syst. Appl., vol. 39, no. 12, pp. 11303–11311, 2012.
- [15] G. E. Vlahos, T. W. Ferratt, and G. Knoepfle, “The use of computer-based information systems by German managers to support decision making,” Inf. Manag., vol. 41, no. 6, pp. 763–779, 2004.
- [16] Z. Gyöngyi and H. Garcíía-Molina, “Web Spam Taxonomy,” in Proceedings of the 1st International Workshop on Adversarial Information Retrieval, 2005.
- [17] <https://prowly.com/magazine/stop-spam-backlinks-ruining-google-reputation/>
- [18] <https://webmasters.googleblog.com/>
- [19] <https://prowly.com/magazine/stop-spam-backlinks-ruining-google-reputation/>