CONCEPTUAL RESEARCH PAPER

COMPUTING TRENDZ Vol VII No. 1,2 , January - December 2017 Print ISSN : 2230-9152, Online ISSN 2456-138X

A study of Quality Content Mining

Abdul Karim Siddiqui
Research Scholar, Computer Applications, Deptt. Of Computer Science & Engineering, LPU- Punjab,

ABSTRACT
Quality Content availability and its accessibility on top searching rank is big issue before us. The paid inclusion of search engine market has made things worse. Data mining is the process of data analysis to discover patterns and relationships in data that may be used to make valid predictions by classification, clustering, prediction, association, extraction and sequence detection of various types of problems. Present Era is the time for searching things of matter (TOM) over Internet and with the advent of World Wide Web (WWW) technologies paradigm, teaching and learning has been shifted towards learners. Today's online environment provides dynamic course materials to learners according to their needs and preferences. There is a need to develop an intelligent computing environment based quality E-learning 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.
Keywords: Content Mining, Things of Matter, TOM, Text Mining, Quality Content, E –Learning, Search Engine Optimization.
Computing Trendz (2017). DOI: 10.21844/cttjetit.v7i1-2.2

Introduction

Today's online environment provides dynamic course materials to learners according to their needs and preferences. In case of E - Content mining, it should be more précised and accurate. It may be termed as "Things of Matter" (TOM) –

The things of online search/ E- Learning contents which matter for you.

Intentional attempts may be made to manipulate search engine rankings by

stuffing specific keywords or keyword phrase queries. This is same to show what

is not right, is right for you. There's a fine line between doing everything you apply to make high rank of your online contents at search engines to that of trying every sneaky trick which is possible.

Corresponding Author: Abdul Karim Siddiqui, Research Scholar, Computer Applications, Deptt. Of Computer Science & Engineering, LPU-Punjab, e-mail: abdul.momentum@hotmail.com
How to cite this article: Siddiqui, A.K. (2017). A study of Quality Content Mining. Computing Trendz 7(1&2): 6-12
Source of support: Nil
Conflict of interest: None

Statement of Research Problem

The keywords we usually type for our content mining on search engines may be polarized towards what these search engines want. This creates a clumsy situation before us to digest what they try to serve! It's our tendency to visit only those contents which are on top ranking or at first search results. The biggest issue in front of quality content Mining is our dependencies on top links and URLs when we search a genuine content.

Relevant literature

- P. Chahal, M. Singh and S. Kumar "Ranking of Web Documents using Semantic

Similarity" : They proposed a better idea to make one's search data quite efficient. The technique gives a relationship or between searched document and user query.

- Parveen Rani and Er. Sukhpreet Singh, "An Offline SEO (Search Engine Optimization) Based Algorithm to Calculate Web Page Rank According to Different Parameters" :

This particular paper shows the new algorithm for calculating web page rank called M-HITS It is a new version of HITS algorithm. In this algorithm six parameters are used to evaluate rank for web page. This technique may be proved fruitful for Future work by using some AI techniques as well.

- Poonam Chahal, Manjeet Singh & Suresh Kumar, "Ranking of Web Documents using Semantic Similarity": They explored relevant relations between the keywords exploring the user's intention and determined their relevance on web with respect to the query provided by the user.

- Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey", we surveyed the research in the area of Web mining, pointed out some confusions regarded the usage of the term Web mining and suggest three Web mining categories.

Scope of study

Generally people find information of their searched keywords from Internet with the help of search engines. Search engines play a vital role in searching these contents by producing specific pages for users according to their query. Web page ranking plays most important role for search engines. It is a technique that ranks the web pages on behalf of factors of qualities and parameters incorporated in search engines. One of the key factors which has been forgotten while design these search engines are Things of matter (TOM). There is a need to optimize web pages or website to make it friendly at Search Engine Optimization Environment. Web search engines like Google and Yahoo are considered as intermediate between user and information repository present at Internetwork Information Ocean. These Search engines use Crawler, spider, and indexer programs on web pages to confirm their visibility at top rank searching.

Relevance of study

Transforming information and experiences into online texts result E- Learning materials. It is a systematic process which also includes knowledge and skills. Now a day everyone is looking for instructional resources of quality contents. These instructional resources need proper screening before deploying on Internet. These online resources should reduce the cost of acquiring right contents and should meet the needs of society. Intelligent e-learning has become our utmost necessity and somehow it matters for us i.e. Things of Matter (TOM).

Search Engine Optimization

Search Engine Optimization is the process of optimizing the rate of searching of any content or website on Internet by applying some SEO tools. A search engine returns the result of related keywords in broad two ways: Organic and paid search results. Organic search results give the list of web pages and URLs, which are the most closed to typed keywords. Our search queries are processed to give most relevant contents as search results. It is also termed as 'natural Search'. The Paid results are shown because of their paid inclusions to Search engines like Google and Yahoo. The owner of

website purchases some specific keywords to gain web searcher's search demands and to show them fix advertisements links at the top or right of search engine results lists.

Data Mining

Data mining is the process of examining large pre-existing databases to generate new information. Web Mining is the discovery and analysis of useful information from the World Wide Web. Web mining is subset of data mining. Web mining is further categorized into two subsets namely web content mining and web usage mining. Content mining focuses on the content of single web page whereas web usage mining captures the identity or origin of Web users along with their browsing behavior at a Web site. Mining of aural and visual data such as sound and video content have also been trending in public demands and that's why they are things of matter.

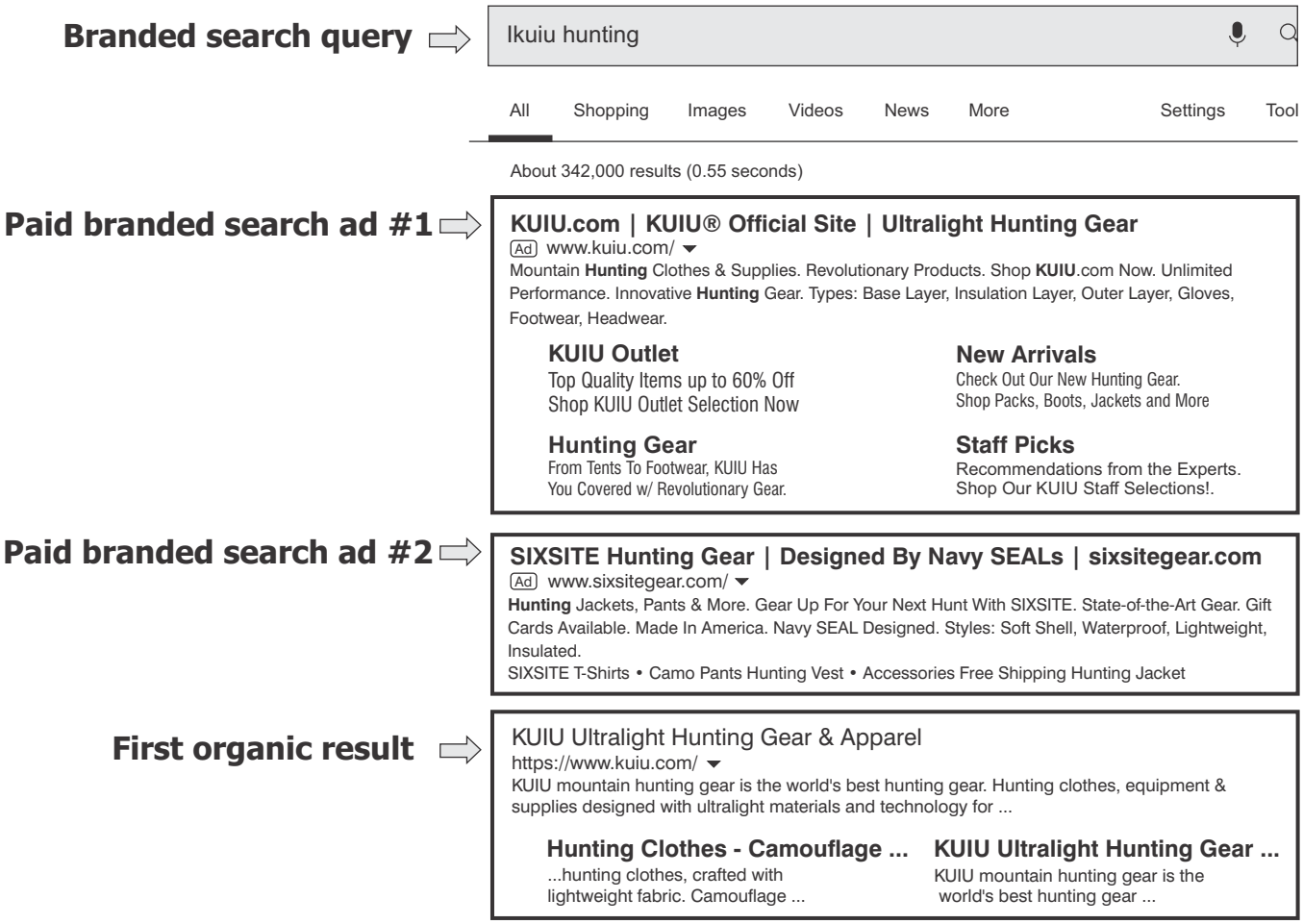

Fig-1 Paid and Organic Search for Branded Item

Objectives

The objective to search quality contents without compromising forcibly flashed E- contents and URL links, has remained a big issue. The searched contents should reflect genuine information in structured and reliable way. Today search is performed by searching the exact Keywords entered by the user. There may be a reason of not knowing exact keywords for many users. A wrong selection or less important contents may divert from right learning. And the objectivity of E-learning may be compromised. There should be an improved model to analyze and promote balanced and quality E- learning text mining. And furthermore, it is not very much confirmed whether you put exact or mixed keywords phrase to search your contents, you will get a genuine and quality E-contents.

Mining – A converging Research Area

Web content mining refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the data available on Web(s). The Web Content mining is a converging research domain from several research communities, such as database (DB), information retrieval (IR) and Artificial Intelligence (AI) especially from machine learning and natural language processing (NLP).

Web mining				
	Web Content Mining		Web Structure Mining	Web Usage Mining
	IR view	Db View		
VIEW OF DATA	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
MAIN DATA	-Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Serves Logs -Browser Logs
REPRESENTATION	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph. -Relational	-Graph	-Relational Table -Graph
METHOD	-Machine learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
APPLICATION CATEGORY	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation andmanagement -Marketing -User Modebng

Table-1 Web mining categories

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 prominent position in search engine result pages by means of paid inclusions. A multi-billion dollar

industry has been evolved to search engine optimization (SEO). The market of SEOs have been growing rapidly. Web spam can be noticed into many forms-

- Keyword stuffing- i.e. populating pages with highly searched or highly monetary reward terms;
- Link spam- i.e. creating cliques of tightly interlinked web pages with the goal of biasing link-based ranking algorithms such as Page Rank(PR)
- Cloaking- i.e. serving substantially different content to web crawlers than to human visitors (Targeting highly trafficked keywords, but presenting completely different content to visitors.)
- Content Stuffing – Adding huge blocks of text in really small font near the footer of a webpage so that the content gets picked up by the search engines, but visitors don't really pay attention to it.
- Buying Links – Offering website owners or webmasters money in exchange for links.
- Widgets – Creating embeddable widgets and tools with links that point back to the widget creator.
- Article / Link Farms – Either building dozens of sites in order to amass links or submitting content to networks of sites that would publish articles that all link back to your site.
- Overly "Linky" Footers – One common practice was to link to high-value webpages from a website's footer.

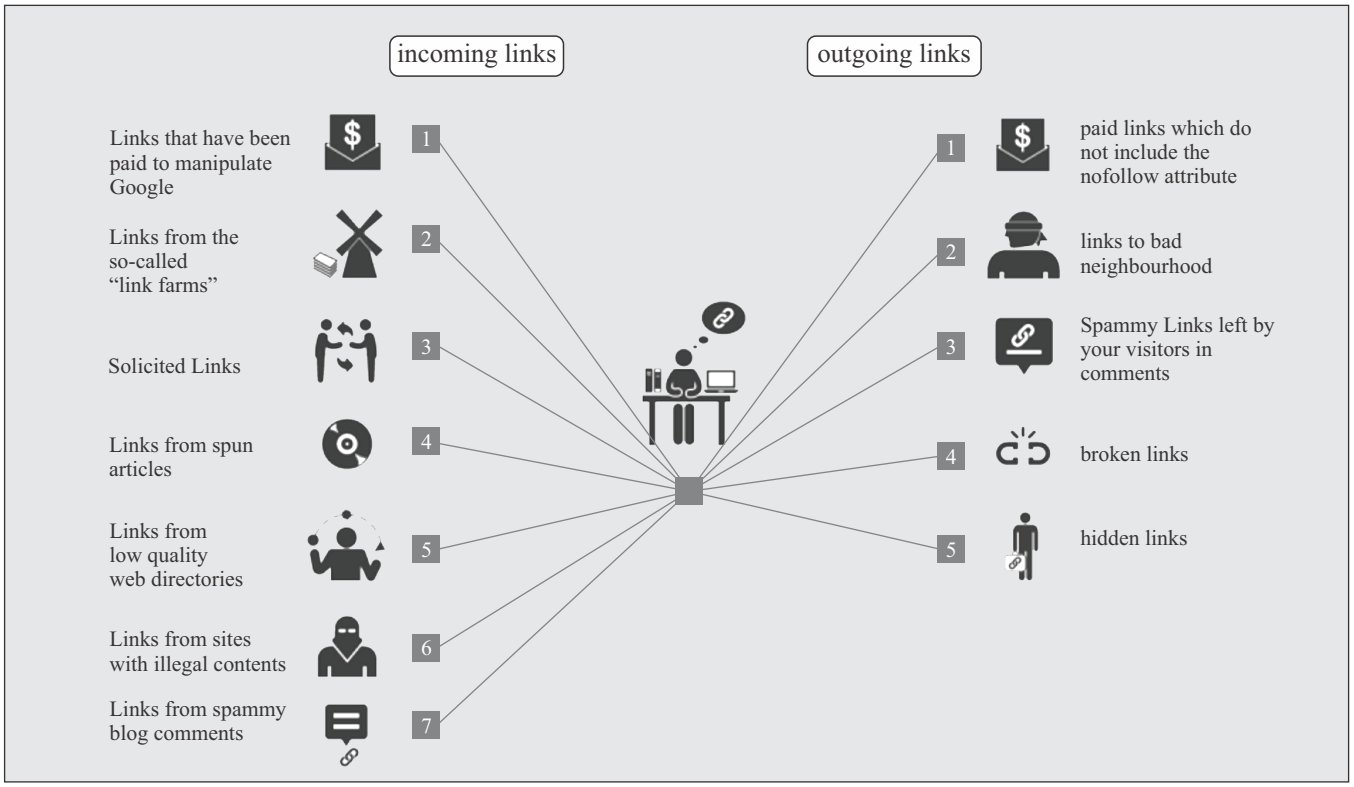

Fig 2 Bad Back links

Methodology

PageRank (PR)

PageRank (PR) is an algorithm used by Google Search Engine to rank web pages in their search results. It is a way of measuring the importance of website pages. A numerical weighting to each element of a hyperlinked set of documents is assigned in page ranking algorithm. The numerical weight, it assigns to given element E is referred by PR(E).

To understand the PR algorithm Let us define an initial value for each be 0.25. The PageRank transferred from a given page to the targets of its outbound links upon the next iteration is divided equally among all outbound links.

If the only links in the system were from pages B, C, and D to A, each link would transfer 0.25 Page Rank to A upon the next iteration, for a total of 0.75.

PR(A) = PR(B) + PR(C) + PR(D)
Suppose instead that page B had a link to pages C and A, page C had a link to page A, and page D had links to all three pages, thus upon the first iteration, page B would transfer half of its existing value, or 0.125, to page A and the other half, or 0.125, to page C. In Same way Page C and D would take place. At the completion of this iteration, page A will have a Page Rank of approximately
0.458.

In Simplified way:-
PR(A) = PR(B) / L(B) + PR(C) / L(C) + PR(D) / L(D)

In the general case, the PageRank value for any page u is

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)}$$

Analysis and Discussions

We find that Page Rank algorithm has a significant role to search any keyword specific content at search engine. The concept was first introduced by founders of Google Search engine. It has been one of the major algorithms to find relevant result online. 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 with semantic search information gathered from a wide variety of sources. It enables to search for things, people and places that Google knows about Landmarks, celebrities, cities, sports teams, buildings, geographical features, celestial objects, works of art, movies and more. You get information instantly which is relevant to you. This is a revolutionary step towards building the next generation of search. The collective intelligence of the webs make your search better than that were in past.

Google's Knowledge Graph contains more than 500 million objects. More than 3.5 billion facts about and relationships between these different objects are pooled.

It enhances Google Search in effective way based on:

- Find the right thing:
- Get the best summary
- Go deeper and broader

Conclusion

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 like E- learning and benefitting common people by availing low cost and quality search information should be in our priorities. These priorities may also be termed as Things of Matter (TOM).

References

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.

Muhammd Jawad Hamid Mughal, " Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, 201.8

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.

Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey", kuleuven.ac.be Volume 2, Issue 1 - page 15

A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in Proceedings of the 15th International World Wide Web Conference, 2006.

Z. Gy¨ongyi and H. Garc´ıa-Molina, "Web Spam Taxonomy," in Proceedings of the 1st International Workshop on Adversarial Information Retrieval, 2005.

https://prowly.com/magazine/stop-spam-backlinks-ruining-google-reputation/

https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph- things-not.html