ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2017.0205

PRESENTING SEARCH RESULT WITH REDUCED UNWANTED WEB ADDRESSES USING FUZZY BASED APPROACH

Nancy Jasmine Goldena
Department of Computer Applications, Sarah Tucker College, India

Abstract

Big Data is now the most talked about research subject. Over the year with the internet and storage space expansions vast swaths of data are available for would be searcher. About a decade ago when a content was searched, due to minimum amount of content often you end up with accurate set of results. But nowadays most of the data, if not all are sometimes vague and not even sometime pertain to area of search it was indented to. Hence here a novel approach is presented to perform data cleaning using a simple but effective fuzzy rule to weed out data that won't produce accurate data.

***Keywords*:**

Big Data, Data Cleaning, Fuzzy Rule Based Approach, Metadata

1. INTRODUCTION

The Internet was initiated in the 1960's as a research work to build fault tolerant and robust communications via computer networks [1]. Over the decades it has grown as not only a communication channel but has transformed into a data sharing hub and repository. The problem with huge amount of data is accessibility. The data are stored in servers which are displayed via a webpage. In order to access the webpage one must know its address. With the amount of websites available it is almost impossible for one to remember all these websites. The other problem is that for a single topic there are multiple websites giving information through varied media. This problem was felt after 1990. To solve this, a software system name web search engine was introduced. Gopher was one well known search engine but after 1996 Google predominantly dominated the search engine world [2].

The search engine basically scans the whole world wide word searching for similar keywords put up in its search query by the user. Once matches are obtained there are presented to the user in a certain predominant order. In order for a search engine to perform at its zenith, it should have the capability to search or retrieve data within the shortest amount of time with high accuracy.

Over the years data housed in the internet has grown steadily and exponentially. Fig.1 shows data growth for the current decade [3]. It could be noted that the data is now measured in zeta bytes. If the projected growth of the Internet is estimated to be 40 zeta bytes it means 40 trillion gigabytes. Our hard drive now comes equipped to store and process more than 1024 GB only and is measured in terabytes. And with the Internet of things increasing there is sure to be more data. The Fig.2 shows the number of devices that will be adding information to the internet [15]. A new field called big data has come up to deal with the issues of huge quantity of data. This has led to searching the right content a herculean task. Search engines implement varies algorithm and techniques to sift through these vast amounts of data to mine the needed information for the user. There are various problems a search engines encounters.

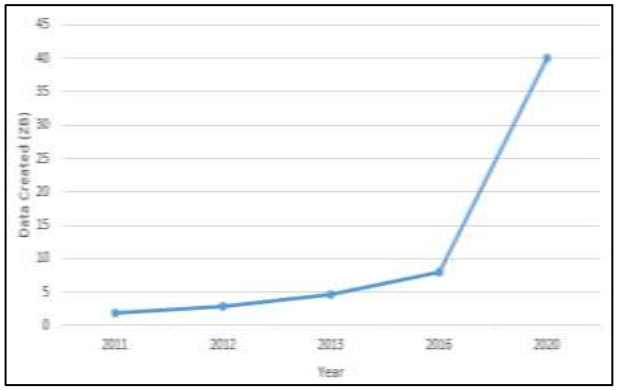

Fig.1. Projection of Data Growth [3]

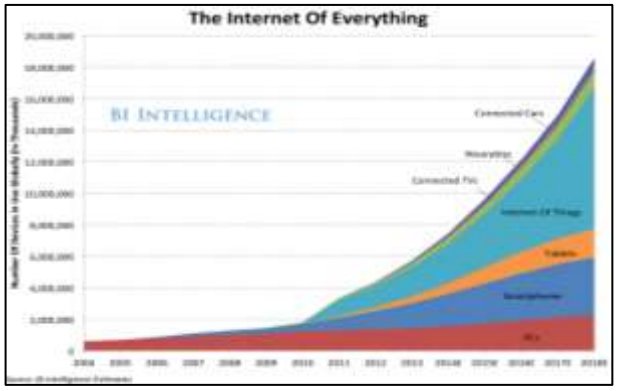

Fig.2. Number of devices that are said to be connected to the internet by 2018 [15]

One problem that search engines faces is to present the most reliable information to the user. A normal search like for example a search phrase like computer would bring up search results about latest computers available in the market rather than about what a computer is that the user may be expecting. The internet search engine has got so coagulated due to vast amounts of data that it is normally hard to get anything specific but it is easy to get loads of data for any random unwanted query.

To solve these underlying problems a novel approach is proposed here, which focuses on exploiting the topic sensitivity part by clustering the text which will be first segmented and them subsequently clustered using the fuzzy based rule approach. The approach is straight forward and very simple. Those having clusters of large size can be representing the duplicate data and can be dealt with as such.

2. LITERATURE SURVEY

Before going into the details of our approach let us see how search engines approach this problem. Most notable search of date is Google. It uses a page rank algorithm by estimating the number of links to the pages. It's functioning is more or less similar to Journal Indexing. It stresses on the concept that more relevant your webpage content is more people would refer to it in their web pages. But sadly this is not the case nowadays. Also people can pay Google to get a better rank to display their product among the top of the search result page [4].

Here in this work, a unique approach is being worked out to get better search results on the page. A number of different heuristics like hashing etc. exist. But here fuzzy ontology is used [5]. Predominantly page ranking comes under the focus of web mining where mostly three techniques are utilized to rank pages namely web content mining, web structure mining, and web usage mining. Ranks are computed to these categories via backlinks, forward links or topic sensitivity [7]. There are various techniques available in the literature to rank a webpage [5-12].

Susan et al. [5] used a technique which ranks pages by what a person sees and what others in a group, like in an institution see. Using this two pronged approach they rank the items of interest of a particular group. This works well for a small group but it is impossible to implement this in a Big Data scenario. This same strategy is again followed by Zhou et al. [6] by giving weightage to various search terms. They do this by identifying similar webpage contents and assign weights accordingly. In A Big Data setup this approach cannot be viable as there are a lot of data and is impractical to give weightage for each and every similar web page. These two methods highlighted here employs weights to rank web pages. There are a lot of other methods reviewed by Ganeshiya et al. [7] and Jindal et al. [8]. Some of the other approaches to rank a page are usage of distributed algorithms, mathematical and statistical models [9-12].

All the above methods discussed in the literature have one thing in common and that is that they all predict what the page rank might be based on various criteria's. The best alternative to this is for the webpage to tell what it is. If that can be achieved then ranking would be less computational accurate and error prone. To do this metadata is used. Metadata are descriptive tags in an Html page, and it contains various information of the webpage [8] [13]. Using metadata the webpage page can describe itself better. We shall discuss what metadata is in detail and how it is incorporated in this work, in the fore coming section in detail.

3. PROPOSED APPROACH

Search Engine Optimization is set of techniques, which are used by the search engine to improve its ranking. Each website contains something called the metadata. Metadata gives certain description about the webpage. Some are being considered for this work are title, Meta Description, Meta Keywords.

Title serves as a heading telling what's in a given web site or page. It is the first entry the crawler encounters. Titles which resemble the search phrase are taken and the rest are left out. Google displays the title as summary in its results page. Meta Description provide a short summary as to what is in the concerning site. Whereas keywords are a bunch of related words that match the content of the webpage or website. Apart from this Meta content there other optional meta tags like Meta Revisit tag, Meta Distribution, Meta Author and Meta Language. But to gain insight into what is in a website the selected above three are sufficient. These meta information can be assessed by crawlers [13].

Fuzzy logic is a technique which employs multi valued logic as opposed to one. The truth value of this technique lies anywhere between 0 and 1 as opposed to the traditional approach of either 0 or 1. The rule determining what should go near 0 or 1 or in other words partial false and partial truth is called the fuzzy rule. As the result is not clear cut it is termed as fuzzy. The rule in question doesn't contain one criteria but many. It is best used in classification [14].

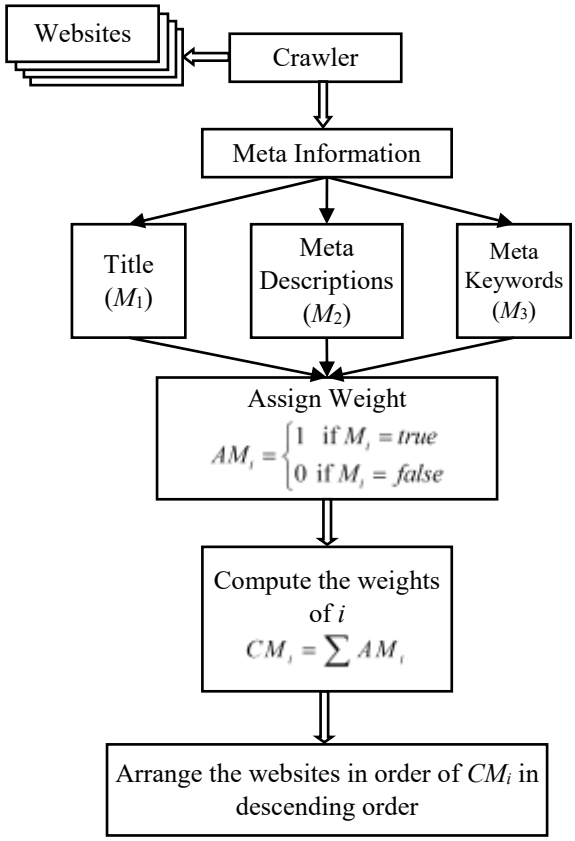

Fig.3. Architecture Diagram

The proposed methodology here extracts the title, Meta description and meta keyword form a website and using a Fuzzy rule based approach which tries to classify the sites into two categories. One category will have site address of contents matching the Keyword and the other the opposite. Data mining employs two very basic approaches to classifying the given data. One approach is the use of clustering where the groups of related data tend to club together and the other approach is classification where certain rules are laid so that different classes can be made. Here the approach is straight forward. Similar patterns of text tend to collaborate or cluster together. The algorithm is as follows

1. Crawl for a website.
2. Extract the meta information viz, title (M_1), Meta Description (M_2) and Meta Keywords (M_3)
3. Analyze the Meta Information
4. Add a weight of 1 for a matching metadata else assign a weight of 0 using the following equation which will be the

fuzzy rule.

$$AM_i = \begin{cases} 1 & \text{if } M_i = true \\ 0 & \text{if } M_i = false \end{cases}$$

5. Sum the weights obtained for each website crawled

$$CM_i = \sum AM_i$$

6. Now arrange the websites in order of the highest weightage.

According to this technique only the one that have three metadata will definitely have the nearest value that is probability wise. When taking into consideration any decision the opinion of the crowd always makes more sense. Here instead of the traditional linkage system, the content in the metadata is exploited. A shopping website will not have any of the meta information. Here, only legitimate sites, sites which do not harbor any threats are taken into consideration.

4. EXPERIMENTAL RESULTS

To understand the working of work a small sample metadata set is analyzed and taken for illustration here as shown in Fig.3. Clearly as you can see the metadata keywords are provided for a few web pages only. Out of the thirty results taken as example here only 8 only have keywords. Keywords generally tell the searcher what content is placed in their site. Hence helping the searcher to get an accurate picture of what is in the website. The websites are ranked as shown in Table.1 where a rank is applied for each field in proportion to its presence, denoted as 0 or 1. These are then summed up and those with a cumulative score. Those that score the perfect 3 as in this case have the accurate data description possible. Hence these can be used for data cleaning. The Fig.4 shows the percentage of reduction in graphical format which shows the hypothesis in greater detail. The accuracy that can be achieved by this method is 100%. In the sample set presented above there are only 10 correct entries pertaining to the mechanical field. The error rate is plotted in Fig.5 and the output is there to see. Of the fifteen that would provide more suitable content crawlers can be employed to know what is in the webpage thereby increasing privacy. Here since we have not programmed a crawler we believe that having these three meta information would give a better picture. An e-market website like eBay won't have these entries and hence can be weeded out.

Input URL	Title	Meta Description	Meta Keywords
store.steampowered.com â€º All Games â€º Massively Multiplayer Games â€º Entropy			
https://aatishb.github.io/entropy/	Entropy Explained	An explorable blog post about entropy, with sheep.	
https://en.wiktionary.org/wiki/entropy	entropy - Wiktionary		
www.forbes.com/sites/quora/2016/.../what-is-an-intuitive-way-to-understand-entropy			
https://www.facebook.com/entropymagazine/	Security Check Required		
www.ruleandmake.com â€º Products			
https://entropyresins.com	Epoxy Resins Glue System | Bio Resin | Epoxy Adhesive	Epoxy bio resins glue by Entropy Resins are useful for many of our activities and hobbies, and the epoxy adhesive can be purchased here!	
https://github.com/buildertools/entropy	GitHub - buildertools/entropy: An entropy and failure injection management API for Docker platforms.	An entropy and failure injection management API for Docker platforms.	
http://entropy.works	Entropy		
https://www.kickstarter.com/.../entropy-thematic-fast-paced-game-of-risk-and-decep			
http://web.mit.edu/16.unified/www/FALL/thermodynamics/notes/node54.html	7.1 Entropy Change in Mixing of Two Ideal Gases	7.1 Entropy Change in Mixing of Two Ideal Gases	notes
http://hyperphysics.phy-astr.gsu.edu/hbase/Therm/entrop.html	Entropy		
https://en.wikipedia.org/wiki/Entropy	Entropy - Wikipedia		

https://www.britannica.com/science/entropy-physics	entropy | Definition and Equation | Britannica.com	How much thermal energy per unit temperature cannot do useful work in a system.	entropy, encyclopedia, encyclopaedia, britannica, article
http://www.nmsea.org/Curriculum/Primer/what_is_entropy.htm	What is Entropy		
https://www.merriam-webster.com/dictionary/entropy	Entropy | Definition of Entropy by Merriam-Webster	Define entropy: a measure of the unavailable energy in a closed thermodynamic system that is also usually considered to be a measure of the system's?	entropy, entropies, entropic, entropically, definition, define, meaning, dictionary, glossary, free, online, english, language, word, words, webster, websters, merriam-webster
https://www.bsigroup.com/en-GB/our-services/business-improvement-software/	Business improvement software | BSI Group	BSI offers a number of software solutions, including Entropy?, to improve your business and help it grow.	(Entropy) Business improvement software
http://www.dictionary.com/browse/entropy	Entropy | Define Entropy at Dictionary.com	Entropy definition, (on a macroscopic scale) a function of thermodynamic variables, as temperature, pressure, or composition, that is a measure of the energy that is not available for work during a thermodynamic process. A closed system evolves toward a state of maximum entropy. See more.	entropy, online dictionary, English dictionary, entropy definition, define entropy, definition of entropy, entropy pronunciation, entropy meaning, entropy origin, entropy examples
https://www.vocabulary.com/dictionary/entropy	entropy - Dictionary Definition : Vocabulary.com	The idea of entropy comes from a principle of thermodynamics dealing with energy. It usually refers to the idea that everything in the universe eventually moves from order to disorder, and entropy is the measurement of that change.	
https://www.entropy.com.au/	Wooden Toys, Educational Toys, Online Toys Australia | Entropy	At Entropy, you will find a broad selection of wooden toys, educational toys, puzzles and kid's toys. Same Day Dispatch. Free shipping and wrap over $125.	Wooden Toys, Educational Toys, Kids Puzzles, Educational Games, Developmental Toys, Baby Toys, Ride On Toys, Toddler Toys, Learning Toys, Activity Toys
http://entropysa.com.au/	Entropy		
http://www.thefreedictionary.com/entropy	Entropy - definition of entropy by The Free Dictionary	Define entropy. Entropy synonyms, entropy pronunciation, entropy translation, English dictionary definition of entropy. n. pl. en?tro?pies 1. Symbol S For a closed thermodynamic system, a quantitative measure of the amount of thermal energy not available to do work. 2.	entropy, online dictionary, thesaurus, dictionary, English dictionary, entropy definition, definition of entropy, legal, medical, encyclopedia, term, law, explanation, information
http://www.entropylaw.com/	Entropy, the first and second laws of thermodynamics and the law of maximum entropy production	The law of entropy, or the second law of thermodynamics, along with the first law of thermodynamics comprise the most fundamental laws of physics. Entropy (the subject of the second law) and energy (the subject of the first law) and their relationship are fundamental to an understanding	entropy, the law of entropy, the second law of thermodynamics, entropy law, 2nd law, disorder, the law of maximum entropy production, Rod Swenson, MEP, LMEP

		not just of physics, but of biology, psychology, and culture. This site contains an introduction, through simple text, links and references, in an easy to understand form to these most fundamental laws as well as the newly recognized “law of maximum entropy production” or MEP and its consequences.	
http://www.mdpi.com/journal/entropy	Entropy | An Open Access Journal from MDPI	Entropy, an international, peer-reviewed Open Access journal.	
http://entropysimple.oxy.edu/content.htm	Entropy Is Simple...If You Avoid The Briar Patches!	Simple introduction to entropy, entropy and nature.	entropy, second law of thermodynamics, entropy and second law of thermodynamics, thermodynamics, disorder, entropy and nature, second law of thermodynamics and nature
http://entropymag.org/	Entropy		
http://www.mdpi.org/entropy/			
http://entropy-project.eu/	Entropy		
https://www.robertsmithson.com/essays/entropy.htm	Robert Smithson	Official web site for the Estate of Robert Smithson, renowned earthwork artist, presenting images and text of earthworks/land art: Spiral Jetty, Amarillo Ramp, land art, nonsites, slide works, conceptual writings and more...	Robert Smithson, robert smithson, smithson, Smithson, estate of robert smithson, Estate of Robert Smithson, nonsites, non-sites, site,displacement, spiral jetty, Spiral Jetty, Amarillo Ramp, amarillo ramp, photoworks, slide works, conceptual writings, conceptual, land art, earth art, earthworks, earthwork, earth works, earth work, asphalt rundown, Asphalt Rundown, Rundown, rundown, swamp, Swamp, James Cohan Gallery, james cohan gallery, james cohan, James Cohan
http://entropymag.org/category/where-to-submit/	Where to Submit? Entropy		

Fig.3. Sample set of website metadata for search phrase entropy

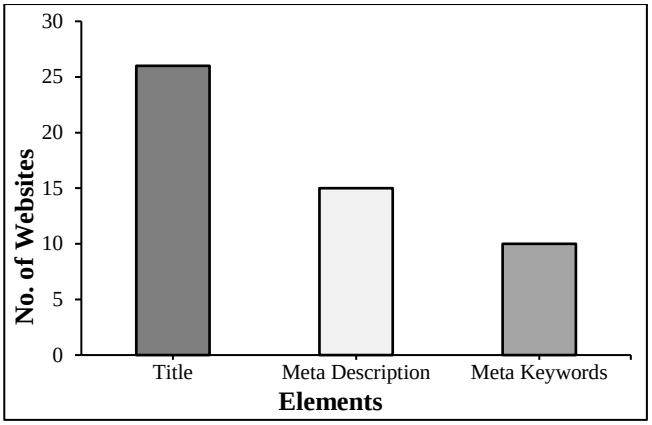

Fig.4. Graphical representation of the Hypothesis

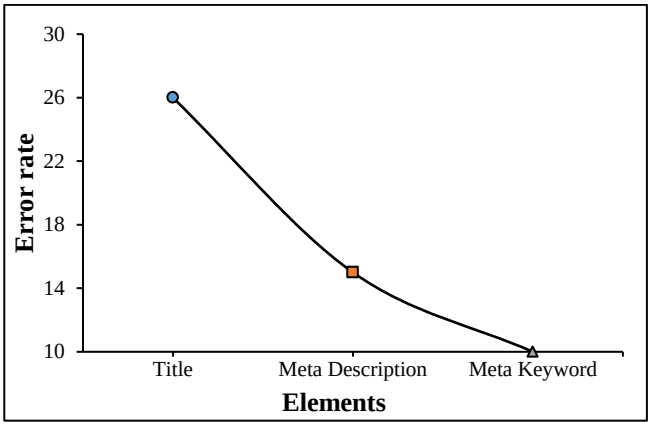

Fig.5. Error rate of the three techniques

Table.1. Comparison of Three Meta Parameters

Input URL	Title	Meta Description	Meta Keywords	Total
1	1	0	0	1
2	1	1	0	2
3	0	0	0	0
4	1	0	0	1
5	0	0	0	0
6	1	0	0	1
7	1	1	0	2
8	1	1	0	2
9	0	0	0	0
10	1	0	0	1
11	1	1	1	3
12	1	0	0	1
13	1	0	0	1
14	1	1	1	3
15	1	0	0	1
16	1	1	1	3
17	1	1	1	3
18	1	1	1	3
19	1	1	0	2
20	1	1	1	3
21	1	0	0	1
22	1	1	1	3
23	1	1	1	3
24	1	1	0	2
25	1	1	1	3
26	0	0	0	0
27	1	0	0	1
28	1	0	0	1
29	1	1	1	3
30	1	0	0	1

Table.2. Comparison of Google Page Rank with the proposed Technique

Input URL	Google Page Rank	Page Rank as per Fuzzy Rule Based Approach
1	1	11
2	2	14
3	3	16
4	4	17
5	5	18
6	6	20
7	7	22
8	8	23
9	9	25
10	10	29
11	11	2
12	12	7
13	13	8
14	14	19
15	15	24
16	16	1
17	17	4
18	18	6
19	19	10
20	20	12
21	21	13
22	22	15
23	23	21
24	24	27
25	25	28
26	26	30
27	27	3
28	28	5
29	29	9
30	30	26

To see if this technique really works well it is pitted against the google page rank technique. The Input URL in Table.1 is the Google Page Rank. According to our fuzzy based approach the Google Page Rank for each page should be as given in Table.2. The amount of variance is depited in a graph format in Fig.6.

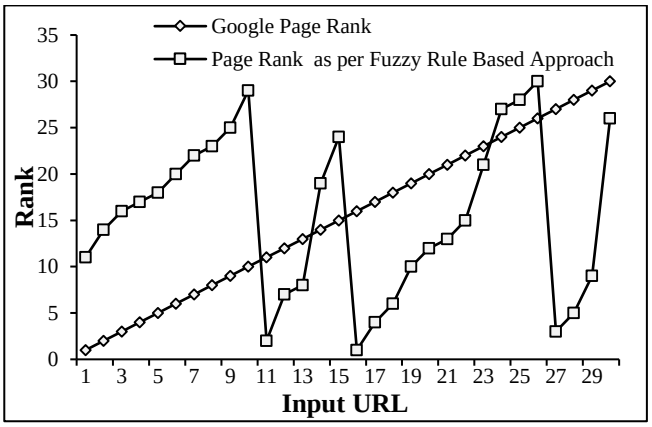

Fig.6. Deviation between the Page Rank of Google and the proposed fuzzy rule based technique

5. CONCLUSION

Hence a sample approach towards refining a search phrase is presented here. The result obtained here weeds out unwanted results and presents with the most accurate results available. In the future especially in the era of Big Data this would help in providing accurate and meaningful results. It also lays a corner stone approach to search the Deep web. Normally most of the websites are ranked by search engines, based on their popularity. This means if you own a webpage and if it is popular you can see your webpage featured in the top ranked page. So an approach like this would definitely make search engine results more versatile. Also in future more such data mining techniques could be experimented to present search engines results apart from popularity and products.

REFERENCES

[1] George Suciu, Alin Geaba, Cristina Butca, Victor Suciu and Octavian Fratu, "Basic Internet Foundation", *Proceedings of International Conference on Future Access Enablers of Ubiquitous and Intelligent Infrastructures*, pp. 255-262, 2015.

[2] Ramesh Manuvinakurike, Wayne F. Velicer and Timothy W. Bickmore, "Automated Indexing of Internet Stories for Health Behavior Change: Weight Loss Attitude Pilot Study", *Journal of Medical Internet Research*, Vol. 16, No. 12, pp. 12-19, 2014.

[3] Big Data and the Internet of Things, Available at: https://www.infragistics.com/community/blogs/mobileman/archive/2015/12/15/big-data-and-the-internet-of-things.aspx.

[4] Facts about Google and Competition, Available at: http://googlecompetition.blogspot.in/.

[5] Susan Dumais, Edward Cutrell, J. J. Cadiz, Gavin Jancke, Raman Sarin and Daniel C. Robbins, "Stuff I've Seen: A System for Personal Information Retrieval and Re-Use", *ACM SIGIR Forum*, Vol. 49, No. 2, pp. 28-35, 2016.

[6] Zhou Hao, Pu Qiumei, Zhang Hong and Sha Zhihao, "An Improved Page Rank Algorithm based on Web Content", *Proceedings of IEEE 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science*, pp. 28-35, 2015.

[7] Deepak Kumar Ganeshiya and Dilip Kumar Sharma, "A Survey: Hyperlink Analysis in Webpage Ranking Algorithms", *Proceedings of IEEE International Conference on Soft Computing Techniques for Engineering and Technology*, pp. 1-8, 2014.

[8] Vikas Jindal, Seema Bawa and Shalini Batra. "A Review of Ranking Approaches for Semantic Search on Web", *Information Processing and Management*, Vol. 50, No. 2, pp. 416-425, 2014.

[9] Atish Das Sarma, Anisur Rahaman Molla, Gopal Pandurangan and Eli Upfal, "Fast Distributed Pagerank Computation", *Theoretical Computer Science*, Vol. 561, pp. 113-121, 2015.

[10] Christian Borgs, Michael Brautbar, Jennifer Chayes and Shang-Hua Teng, "Multiscale Matrix Sampling and Sublinear-Time Pagerank Computation", *Internet Mathematics*, Vol. 10, No. 1-2, pp. 20-48, 2014.

[11] Jun Yu, Yong Rui and Dacheng Tao, "Click Prediction for Web Image Reranking using Multimodal Sparse Coding", *IEEE Transactions on Image Processing*, Vol. 23, No. 5, pp. 2019-2032, 2014.

[12] Zhijun Yan, Meiming, Dongsong Zhangxing and Baizhang Ma, "EXPRS: An Extended Pagerank Method for Product Feature Extraction from Online Consumer Reviews", *Information and Management*, Vol. 52, No. 7, pp. 850-858, 2015.

[13] Rashmi Janbandhu, Prashant Dahiwale and M.M. Raghuwanshi, "Analysis of Web Crawling Algorithms", *International Journal on Recent and Innovation Trends in Computing and Communication*, Vol. 2, No. 3, pp. 488-492, 2014.

[14] Patricia Melin and Oscar Castillo, "A Review on Type-2 Fuzzy Logic Applications in Clustering, Classification and Pattern Recognition", *Applied Soft Computing*, Vol. 21, pp. 568-577, 2014.

[15] Gartner, IDC, Strategy Analytics, Machina Research, company filings, BI Intelligence Estimates. Accessed on May 2017.