I.J. Information Technology and Computer Science, **2015, 03, 12-17**
Published Online February 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2015.03.02

Credible Mechanism for More Reliable Search Engine Results

Mohammed Abdel Razek[1,2]
[1]Research and Development Department, Deanship of Distance Learning, King Abdulaziz University, P.O. Box 80254, Kingdom of Saudi Arabia
[2]Math. & Computer Science Department, Faculty of Science, Azhar University, Nasr City, 11884, Cairo, Egypt
E-mail: abdelram@azhar.edu.eg

***Abstract*—** the number of websites on the Internet is growing randomly, thanks to HTML language. Consequently, a diversity of information is available on the Web, however, sometimes the content of it may be neither valuable nor trusted. This leads to a problem of a credibility of the existing information on these Websites. This paper investigates aspects affecting on the Websites credibility and then uses them along with dominant meaning of the query for improving information retrieval capabilities and to effectively manage contents. It presents a design and development of a credible mechanism that searches Web search engine and then ranks sites according to its reliability. Our experiments show that the credibility terms on the Websites can affect the ranking of the Web search engine and greatly improves retrieval effectiveness.

***Index Terms*—** Web Search Engine, Credible Web Search, Dominant Meaning, Top-Level Domains

I. Introduction

Nowadays, a huge type of information and knowledge has been dispersed on the Web such as health management and governmental policy, newspapers, weather, time tables, ticket purchasing, and net banking [1]. Web information is not essentially of equal advantage, where definite information seems to be more valuable than other. Therefore, the Web credibility is a crucial aspect in getting the complete ability of the Web. The challenge now is to find a way to distinguish credible Website from trustable Websites. Since long time ago, researchers have been looking for the credible information, so far there is no pure definition of credibility. Rieh [2] defines credibility as "*an intuitive and complex concept that has two key dimensions: trustworthiness and expertise*". Credibility differs from information quality, where this paper focuses on Web credibility which is defined as trustworthiness of a web site [3].

Web search engines play an important role to achieve user satisfaction, however, explicit measures of credibility are not applied [4]. Ordinary web search engines cannot evaluate the content [5]. Using the Internet without search engines to find specific information is like wandering aimlessly in the ocean and trying to catch a specific fish [6], [7]. Lewandowski [8] found that search engines has no fully integrated credibility frame work. Mandl [9] found that most of the technical means for finding suitable indicators for credible Web pages are depending on human conclusions. However, Rieh [10] sees that the credibility is determined based on the people's assessment of whether information is reliable or not, Google search engines rate documents based on Ranking algorithm rather credibility

What this research examines is aspects which are affecting on the credibility of the Websites and how to apply them on a Web search engine to get believable Websites. Many studies have been done to assign these aspects such as [5], [10], and [11]. This research studies and defines suitable aspects which are affecting on the Webpages credibility. To improve the results, we use the dominant meaning technique [7]. This technique considers the original query as a master word and its dominant meaning as slave words.

To be effective, a good representation for the meaning of the domain knowledge of the query helps on retrieving good results. This paper uses a tree to build a relation between the master and its slaves. To figure out the closeness between the master word and its slave words in documents, we use the dominant meaning probability [7]. The re-ranking algorithm includes these aspects will be applied on the research results coming from Web search engines, and then they can be used to re-rank search results to filter a credible Websites. The paper proposes a calculation method of credibility values for Google results. Using this assumption, the credibility algorithm calculates the degree that is to be used in filtering the results. Our proposed procedure consists of the following steps:

1) Send a query to Google,
2) Retrieve information from Google,
3) Extract credibility aspects from each document,
4) Calculate credibility values of each document,
5) Extract slaves words related to the query,
6) Calculate dominant meaning probability of each document,
7) Compute a ranking value for each document,
8) Re-rank retrieved information based on the ranking value,
9) Post ranked information to a user.

This paper is organized as follows: section 2 discusses credibility terms to be used in re-ranking Web search results. Section 3 presents the mechanism of the credibility, the methodology to extract credibility aspects, and the algorithm for re-ranking the results. Section 4

demonstrates experimental results, and finally, section 5 concludes the paper.

II. Credibility Terms

Credibility is investigated by many researchers using definitions, approaches, and presuppositions that are field specific [12]. This section explores the use of aspects to improve web search.

A. Credibility aspects

Credibility can be defined as "*the believability of a source or message, which is made up of two primary dimensions: trustworthiness and expertise*" [12]. Credible people are believable people. Fogg defines four types for Web credibility [3]: presumed credibility where persons recommend Website, reputed credibility which is depended on the reputation of a person included in this information, surface credibility where credibility is given based on the structure of the Webpages, and is assigned based on the expensed of the recommender. However, these credibility terms cannot be miserable, accordingly, they cannot be suitable aspects for search engine algorithm.

Lewandowski [8] summarizes how to apply credibility criteria in search engines in three techniques: Webpages of low credibility are expected from the search engines' indices, marked in the results presentation, or ranked lower in the results lists. The limitation in Web search engine was based on two facts: the algorithm of the indexing has only low barriers for documents, and the same algorithms matched for all Webpages which offer the same chance for retrieving them.

In contrast, this research is looking for suitable aspects which affecting on the credibility of Web sites. Afterward they can be used to re-rank search results coming from Web search engine to filter a credible Websites. Such aspects are known as the top-level domains (TLDs). TLDs is the latest portion of the domain name. ".com, .org or .edu" are some examples of TLDs. TLDs give us more information about the Website, where ".ac and .edu "are represented educational sites, ".com and .biz" are characterized commercial sites, and ".gov" signifies U.S. governmental sites [3]. There are three domain names which can be registered (.com, .net, and .org) without restriction, and the other four (edu, .gov, .int, .mil) have limited purposes. Some other aspects effects on the credibility of the Website such as clear contact information, recently updated Websites, and Websites without ads [3].

B. Problem Statement

Fig. 1 shows results for the query "Barack Obama". Based on the aspects mentioned in last subsection, we can see that the most credible document is number four in the research list "President Barack Obama, the White House", where its top-level domain is ".gov" and its URL includes one slave dominant word "White House" related to the master word "Barack Obama".

Fig. 1. Google results for the query "Barak Obama"

III. Credible Mechanism

Actually, we can get creditable relevant search results if the search query takes into consideration the credibility aspects. The first stage is to send the query to Google, retrieving a list of the Web pages, showing them back to the end-users. Those pages are described as snippets. Those snippets contain a summary the Webpage as a few lines of text appeared under every search result. For each result in the list, the default result page includes [13]: a title, a snippet, a link URL, file size, date, and a link to a cached page. They are parallel analysis: the credibility aspect extraction and dominant meaning extractions.

The credible mechanism gains the benefit of two knowledge bases: URL knowledge bases, and dominant meaning knowledge bases. However, the two bases are used on the time of the query, they are constructed off-line.

A. URL Knowledge Base

Uniform Resource Locator (URL) knowledge base consists of information selected to be used in the credibility algorithm. The credibility aspects applies on a URL, snippet's date and file size. URL is built to replace the IP addresses used in communicating with the server where there the relation between them is one-to-one [14]. There is a server machine for Domain Name System (DNS) that is in charge of translating human-friendly URL into IP addresses.

Fig. 2. Anatomy of a URL

In other words, URL is the web address of an online resource, i.e. a web site or document. Fig. 2 presents the anatomy of a URL based on MOZ Company [14].

As mentioned in section 2, there is a degree of credibility for each TLDs. There are a lot of fraud Websites which causes many problems for Web users. For example, the Website "www.whitehouse.net" is a fake which obviously has the look and feel of an official U.S. government "www.whitehouse.gov". Accordingly, we suggest a table values for the TLDs, which assign high values for the TLDs of government organizations such as ".gov, .eg, .sa". Consequently, the credibility algorithm looks at Top-level domain and use weights attached for each one as in Table 1. Where β is suggested real value and $0 < \beta < 1$.

Therefore, the Top-Level value $Top - Level("D_k")$ is computed based on the Table 1. The second aspect is snippet's date. We think that the more updated date is, the more credible the website is. Suppose that the format of the date is (*dd, mm, yyyy*), then we normalize it by dividing the value by "2043" which means the maximum of the year is "2020", the maximum of the month is "12", and the day is "31". Accordingly, the date value is computed as follows:

$$date(dd, mm, yyyy \mid D_k) = \frac{dd + mm + yyyy}{31 + 12 + 2020} = \frac{dd + mm + yyyy}{2043} \quad (1)$$

The third aspect is snippet's file size. Suppose that $Size("D_k")$ represent the size of D_k . Therefore, the size value of each document is computed as follows:

$$SizeValue("D_k") = \frac{Size("D_k")}{\sum_{k=1}^{M} Size("D_k")} \quad (2)$$

Consequently, the URL value is evaluated as:

Table 1. Top-Level Domain suggested weights

Top-Level Domain	Suggested Weight	Meaning of Top-level Domain
".mil"	0.90 β	U.S. military or affiliated agency
".gov"	0.85 β	governmental agency
".edu"	0.82 β	educational institution
".int"	0.80 β	international organizations
".com"	0.55 β	commercial business
".net"	0.45 β	large network
".org"	0.40 β	nonprofit organization

$$URL("D_k") = \frac{1}{3}[Top - Level("D_k") + SizeValue("D_k") + date(dd, mm, yyyy \mid D_k)] \quad (3)$$

B. Dominant meaning knowledge base

The construction of the dominant meaning knowledge base is done off-line and used online. We follow the same graph definition of Razek [6] called the Dominant Meaning Graph (DMG). This illustrates the relation between the master words and its slaves. In DMG, words is represented as nodes and its relations to others is represented by edges which define the weight between the master and its slaves. Suppose that w_i represents a master word, then we could consider the set $\{w_1, w_2, ..., w_n\}$ represents its slaves if there is a non-negative weight P_{ji} called dominant meaning distance, where $P_{ji} = P(w_j \mid w_i)$ represents how much w_j linked to the word w_i, as shown in Fig. 3.

As shown in Fig. 3, following DMG [13], we can see that the graph presents all edges between nodes using dominant meaning distance value $P(w_j \mid w_i), \forall j, i > 0$.

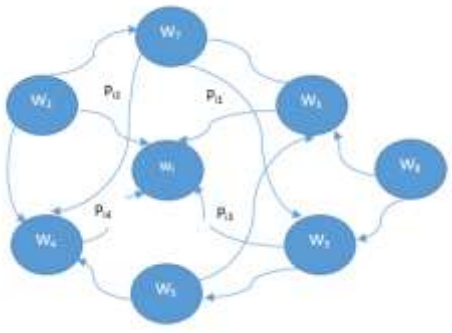

Fig. 3. Dominant Meaning Graph

Dominant meaning space between two words w_i and w_j is calculated as follows:

$$P(w_j \mid w_i) = \frac{1}{N} \left[\sum_{k=1}^{N} \Gamma(w_i, w_j \mid D_k) \left[\frac{(F(w_i \mid D_k) + F(w_j \mid D_k))}{F(w_i \mid D_k)} \right] \right] \quad (4)$$

$$\Gamma(w_i, w_j \mid D_k) = \begin{cases} 0 & \{w_i, w_j\} \cap D_k = \Phi \\ 1 & \{w_i, w_j\} \cap D_k \neq \Phi \end{cases}$$

Where,

$$F(w_i \mid D_k) > F(w_j \mid D_k)$$

Where the functions $F(w_i \mid D_k)$ and $F(w_j \mid D_k)$ represent the frequency of occurrence of the two words w_i and w_j in the document D_k.

Where $\{w_i, w_j\} \cap D_k = \Phi$ means that the two words w_i and w_j do not occur together in the document D_k and $\{w_i, w_j\} \cap D_k \neq \Phi$ means that the two words w_i and w_j occur together in the document D_k.

One of the major challenges of the dominant meaning graph is to determine which words has to be added for re-ranking Google search results. The dominant meaning algorithm is designed to return ten suitable words. In this algorithm, Breadth First Search (BFS) is a suitable way to traverse graphs [15]. It sorts the values $\{P_{i1},...,P_{in}\}$ in decreasing order and then returns the first ten words to be used in the filter. To do so, we have first to build the dominant meaning graph and then create an algorithm to traverse it.

Dominant Meaning Algorithm (Requested word w_i)

1. ***Construct Search* List.**
2. ***While* Search List ≠ [ ] do begin**
   - **I. Put X = the first word in the search list**
     1. **For each $j = 1,...,M$ do,**
        - **w_j = parent (X),**
        - **Compute $P_{ji} = P(w_j \mid w_i)$**
        - **If P_{ji} > 0; then return P_{ji}.**
        - **Else return I ;**
   - **II. Sorting the values of $P_{i,j}$ as decreasing order.**

Return the list $\{P_{i1}, P_{i2},..., P_{iM}\}$.

C. Methods of Re-ranking Results based on credible value

The proposed probability increases retrieval efficiency by striking some limitations on retrieved documents. We suppose that the query with its slaves is $\{w_1, w_2,..., w_n\}$, and the stream of snippets coming from Google search is $\{D_k\}_1^M$. Based on anatomy of a URL and Google search results, we can see that D_k consists of snippet, title, domain, folder, and page. Accordingly, the relevance of document D_k is a numerical value that is intended to reflect how important a document D_k and is computed as follows:

$$\mathrm{P}(w_1,...,w_n \mid D_k) = \frac{1}{n}\left[\sum_{j=1}^{n} \frac{F(w_j \mid D_k)}{F_{\max}}\right] \quad (5)$$

$$F_{\max} = \underset{j=1,...,n}{Max}\{F(w_j \mid D_k)\} \qquad \forall k$$

Function $F(w_j \mid D_k)$ represents the number of occurrence of the word w_j in document D_k.

Following (3), and (5), we can suppose that credible value for each document D_k is computed as follows:

$$Credible(D_k) = \frac{1}{2}\left[\begin{array}{l}\mathrm{P}(w_1,...,w_n \mid D_k) \\ +Credibilty("D_k")\end{array}\right] \quad (6)$$

For all k.

The following section presents the experiment and its results.

IV. Experiments and Results

Our proposed algorithm was run on a commodity PC with Window 8, CORE i7 CPU, 2.4 GHz and RAM 4 GB.

The dataset in our experiments was collected using five queries to a particular search engine, Google, retrieved around top 500 snippets. Some of our colleagues helped us for manually evaluating the dataset and investigate which the Webpage is credible or not credible according to each query. We used 25% of the snippets in each query as a training set for building the DMG, as mentioned in section 3. We then computed $P_{ji} = P(w_j \mid w_i)$ between word in snippets and its query using (4). Therefore, the weight for between each word and the master (query) in the snippets is assigned and used to draw the graph. We implemented the dominant meaning algorithm using Java language and was used to help in compute the value of Dominant meaning space. Table 2 presents the structure of the dataset: the number of snippets retrieved, the number of credible, incredible, and not applicable snippets.

Table 2 Collection used for experiment and tested by humans

Query	Number of Snippets	Number of Snippets		
		Credible	Incredible	Not Applicable
Barack Obama	500	345	85	70
Hilary Clinton	500	360	105	35
Bill Clinton	500	370	100	30
Vladimir Putin	500	355	95	50
Angela Merkel	500	380	90	30

The experimentation conducted on the top 100 snippets of Google search for the same query. Precision and recall are often used to evaluate the efficiency of information retrieval systems. The precision measures the capability of the algorithm to retrieve only relevant items if exist. The recall measures the capability of the algorithm to retrieve all relevant items if exist. We actually evaluated the precision and recall to the top 100 snippets generated

by Google and the proposed algorithm for each of the five queries.

A. The impact of using credible aspects ranking vs. Google ranking

This experiment illustrates the effectiveness of the credible aspects in Google results. Table 1 shows the number of Snippets, the credible items with Google, and credible items with URL knowledge base.

Table 2 shows performance improvement when the results are re-ranked using URL knowledge base. As shown, the best improvement happened in the query number five "Angela Merkel". The lowest improvement is the result of the query "Barack Obama". . Fig. 4 shows the improvement in the precision and recall for Top 100 snippets retrieved using Google with URL improvement.

Table 3. the credible snippets in the top 100 pages using URL aspects vs. Google

Query	Number of Snippets	Number of Snippets	
		Credible items With Google	Credible items With URL Knowledge base
Barack Obama	100	65%	85%
Hilary Clinton	100	71%	83%
Bill Clinton	100	78%	81%
Vladimir Putin	100	80%	87%
Angela Merkel	100	77%	89%

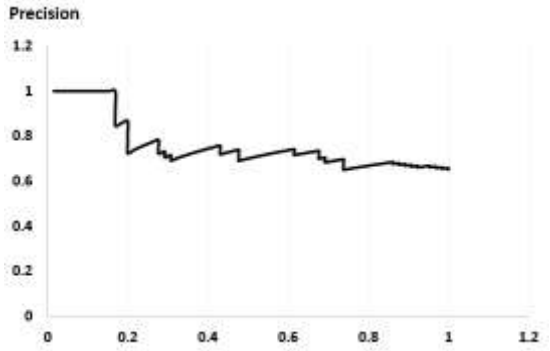

Fig. 4. Precision and recall for Top 100 snippets retrieved using Google

Table 4 presents the improvement using URL knowledge base approach: 11% in credible snippets retrieved

As shown in Fig. 5, the results re-ranked which uses knowledge base produce both a significant boost of precision and recall. It obviously, there is an increasing in the number of relevant top-level snippets as shown in Table 4.

B. The impact of using credible value ranking vs. Google ranking

In this subsection, we organize another experiment for certifying the impact of using dominant meaning along with URL aspects vs. the original query. We compared dominant meaning performance with Google results. Table 4 demonstrates the improvement happened on the results after re-ranking using credible value. As shown, the best improvement happened in the query "Angela Merkel" with 97% improvement. The lowest improvement was in is the query "Vladimir Putin". In general, the improvement was based on the credible value: 21% in credible snippets retrieved.

Table 4. The impact of using credible value ranking vs. Google ranking

Query	Number of Snippets	Number of Snippets	
		Credible items With Google	Credible items With URL knowledge base
Barack Obama	100	65%	95%
Hilary Clinton	100	71%	94%
Bill Clinton	100	78%	96%
Vladimir Putin	100	80%	93%
Angela Merkel	100	77%	97%

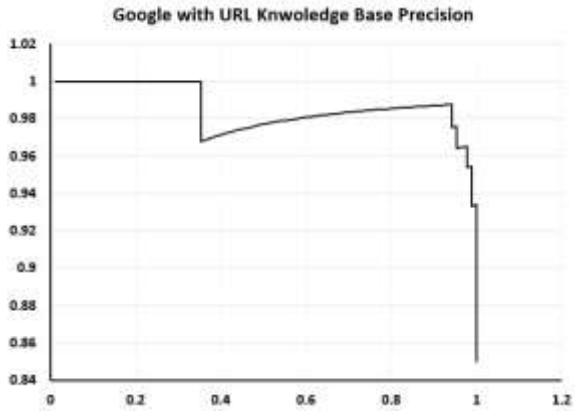

Fig. 5. Precision and recall for Top 100 snippets retrieved using Google with URL improvement

In Fig.6, we merge recall and precision into a single overall measure to clarify the impact of using credible value for ranking Google results vs. Google ranking. It demonstrates the average of precision at each point of recall which is recounted as the summary result for the query "Barack Obama".

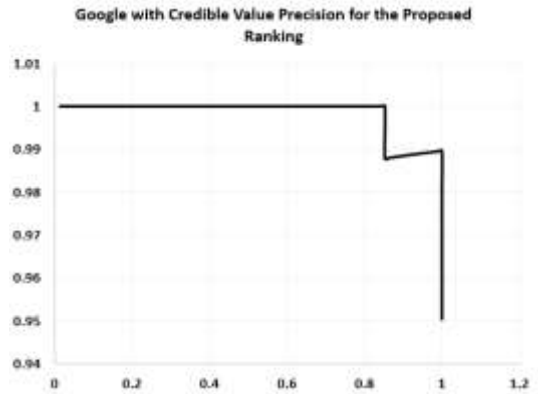

Fig. 6. the impact of using credible value for ranking Google results

In general, Fig. 6 shows that a significant improvement in precision and recall can be improved by re-ranking results using a dominant meaning and URL aspects.

V. Conclusion

This paper presented a novel type of credible Web search engines. This technique used some two credible aspects to improve the Google search results. These aspects include: URL knowledge bases, and dominant meaning knowledge bases. The paper presents a dominant meaning graph to represent the meaning knowledge base and suggested an algorithm to traverse it. We investigated the impact of the proposed credible aspects on re-ranking of the Google search results. Experimental results indicate that the impact generated by our technique is consistently better than that of those results coming from Google. The experiment measured the impact of using credible aspects ranking vs. Google ranking and the impact of using credible value ranking vs. Google ranking. In general, the improvement based on the credible value is 21% in credible snippets retrieved.

References

[1] K. Daisuke, K. Sadao and I., Kentaro (2008) " Grasping Major Statements and their Contradictions Toward Information Credibility Analysis of Web Contents", In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI'08), short paper, pp.393-397, 2008.

[2] B., Hilligoss, & S. Y. Rieh, (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. Information Processing & Management, 44(4), 1467-1484.

[3] B.J. Fogg, Persuasive Technologies: Using Computers to Change What We Think and Do. San Francisco: Morgan Kaufmann Publishers, 2003.

[4] B.J. Fogg, Jonathan Marshall, Tami Kameda, Joshua Solomon, Akshay Rangnekar, John Boyd, & Bonny Brown "Web Credibility Research: A Method for Online Experiments and Some Early Study Results" Proceedings of ACM CHI 2001 Conference on Human Factors in Computing Systems, v.2. New York: ACM Press.

[5] Y., Kammerer, & P., Gerjets, (2012). How search engine users evaluate and select Web search results: The impact of the search engine interface on credibility assessments. In D. Lewandowski (Ed.), Web Search Engine Research (pp. 251-279).

[6] M. Razek, C. Frasson C., M. Kaltenbach, (2006), Dominant Meaning Approach towards Individualized Web Search for Learning. Environments Advances in Web-Based Education: Personalized Learning Environments edited by George D. Magoulas and Sherry Y. Chen © 2006, Idea Group Inc.

[7] M. Razek, Towards More Efficient Image Web Search. Intelligent Information Management 5.6 (2013).

[8] D., Lewandowski (2012) Credibility in Web Search Engines, Apostel, Shawn; Folk, Moe (eds.): Online Credibility and Digital Ethos: Evaluating Computer-Mediated Communication. Hershey, PA: IGI Global, 2012.

[9] T. Mandl, (2006). Implementation and evaluation of a quality-based search engine. Proceedings of the seventeenth conference on Hypertext and hypermedia (pp. 73–84). New York: ACM.

[10] S. Y., Rieh, (2010). Credibility and cognitive authority of information. Encyclopedia of Library and Information Sciences, Third Edition. Taylor & Francis.

[11] B., Hilligoss, & S. Y. Rieh, (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. Information Processing & Management, 44(4), 1467-1484.

[12] A. Flannigan, and M. Metzger (2007), "Introduction, in MacArthur 2007", in Metzger, M. and Flanagin, A. (Eds), Digital Media, Youth, and Credibility, MacArthur Foundation Series on Digital Media and Learning, Chicago.

[13] Google, (2012) "Google Search Appliance Creating the Search Experience" Google Search Appliance software version 7.0, September 2012.

[14] Moz, 2014 "Anatomy of a URL", http://moz.com/learn/seo/url, [last accessed 28 of septamber, 2014]

[15] S.J., Russel, & P.Norvig, (2009). Artificial Intelligence: A Modern Approach. 2nd Ed. Upper Saddle River, NJ: Prentice Hall.

Author's Profiles

Mohammed Abdel Razek holds a Ph.D. in Computer Science - Artificial Intelligence - from the University of Montreal, Canada, in 2004. He has been finished a postdoctoral with Prof. Claude Frasson, PARI project supported by NSERC, Canada. He has more than 40 publication of applying artificial intelligence techniques on e-learning, e-commerce, digital library, and others. In 2006, during his work at Faculty of Science Azhar University, Egypt, he founded the information systems and networks unit. In 2007, he was one of the team work who established 17 e-learning centers at 17 Egyptian universities during his work at National E-learning Center (NELC). By the end of 2007, Dr. Razek was joined the National Authority for Quality Assurance and Accreditation of Education (NAQAAE) and he shared a teamwork to write the e-learning standards. Now, he is working as a consultant for e-learning quality assurance at the deanship of distance learning and a consultant for the Vice President of Development of King Abdulazize University.

How to cite this paper: Mohammed Abdel Razek,"Credible Mechanism for More Reliable Search Engine Results", International Journal of Information Technology and Computer Science(IJITCS), vol.7, no.3, pp.12-17, 2015. DOI: 10.5815/ijitcs.2015.03.02