Architecture for Checking Trustworthiness of Websites

Sana Ansari
Asst. Professor
Don Bosco Institute of Technology, Mumbai

Jayant Gadge
Asst. Professor
Thadomal Shahni Engineering College, Mumbai

ABSTRACT
Today information retrieval from Internet is becoming a commonplace phenomenon. Since, information is readily available and accessible to everyone. Whenever any user types a query in search engine, they would get answers within few micro seconds. However, the results which they get may or may not be accurate because different websites may give different information about the same entity. So, the biggest question is, which website should the user trust?

There are many characteristics using that users can determine trustworthiness in content provided by Web information sources. In the proposed system, the filtering of website trustworthiness is based on five major areas as Authority, Related resources, Popularity, Age and Recommendation. The proposed system defines eighteen factors which are categorized under the mentioned five major areas. The website trustworthiness is calculated based on these eighteen factors of each URL and it is stored thereby increasing the performance in retrieving the trustworthy websites. The objective of the proposed system is to provide more trustworthy websites as top results which would save considerable amount of searching time.

General Terms
Trustworthy, URLs, Information Retrieval, Links, Factors, Webpage, Algorithm

Keywords
Authority, Popularity, Recommendation, Page Rank, Inbound Link, Alexa Rank, WOT, Dmoz Listing

1. INTRODUCTION
Every day, people retrieve all kinds of information from the Web. However, there is no guarantee for the correctness of information as they come from different sources, varying in quality. At times people may get conflicting information from different websites [1]. In current scenario, Internet is the most popular as well as important source of information. If somebody is searching for any kind of information, they would write a query in any search engine (such as google.com, ask.com etc.) and get their answer in few micro seconds. However, the answer, which the search engine provides, may or may not be trustable, because various websites may provide different result for the same query.

For example: If you type in Google following query: (The query was sent on 15 August 2010) "What is the depth of Indian Ocean?" You will find following results:

1. www.eoearth.org gives "3900m"
2. en.wikipedia.org/wiki/Indian_Ocean gives "3890m"
3. www.infoplease.com gives "3400m" and so on….

So, which website should the user rely on? Thus it is clear that the search results are not correlative and credible [3]. To achieve desire data on web, users needs to go through each website manually. Since this process is not only time consuming but also inefficient.

2. LITERATURE REVIEW
The usefulness of a search engine depends on the relevance of the result set it gives back. While there may be millions of web pages that include a particular word or phrase, some pages may be more relevant, popular, or authoritative than others. Most search engines employ methods to rank the results to provide the "best" results first. How a search engine decides which pages are the best matches, and what order the results should be shown in, varies widely from one engine to another. Google, a search engine with a full text and hyperlink database, is designed to crawl and index the Web efficiently and return much more satisfying search results than existing systems. It makes use of the link structure of the Web to calculate a quality ranking for each Web. The rank algorithm used by Google is PageRank. PageRank extends the idea that the importance of an academic publication can be evaluated by its citations to pages on the Web, which can be similarly be evaluated by counting back links. In PageRank algorithm of Google, the ranking value PR of a page A is measured using given formula:

$$PR(A)=(1-d)+d\ (PR(T1)/C(A)+ \ldots + PR(Tn)/C(A)) \quad (1)$$

Where T1…Tn are pages pointing to page A, hence representing backlinks. The parameter d is a damping factor which is scaled between 0 and 1, and C(A) is the number of links leaving page A, hence representing outgoing links. The rank of page A or PR(A) can be calculated using a simple iterative algorithm. As shown by formula (1), the ranking process recursively defines the relevance of page A to be the weighted sum of its backlinks [26].The Page Rank algorithm is used to find out the pages with high authorities. In next, the Hyperlink Induced Topic Search(HITS) algorithm, the first step is to retrieve the set of results to the search query. The computation is performed only on this result set, not across all Web pages. Authority and hub values are defined in terms of one another in a mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values of the pages it points to [8]. These both approaches are identifying most important web pages, but the popularity of web pages do not necessarily lead to accuracy of information. Even the most popular website may contain many errors [6, 7]. Related work has been also done with TRUTHFINDER algorithm. This algorithm is used to find out the trustworthy websites. But there are certain limitations. The first limitation is, initial assumption of Website Trustworthiness is taken as

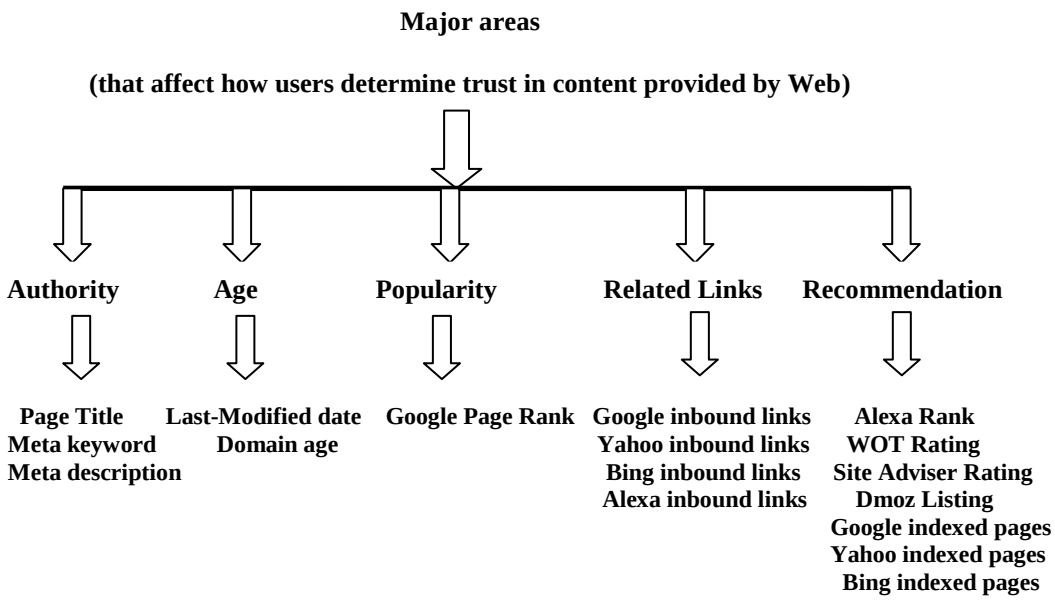

Figure 1: Tree Diagram representing 18 Factors

0.9 in all cases even if it is popular, authoritative or untrustworthy websites. And second limitation is, recalculation of trustworthiness of websites for each query given by the user reduces the performance of the system [1].

3. FACTORS AFFECTING TRUST

While determining trust, it is not a good practice to assume that Web page's main site is the only factor. Apart from this, there are many parameters that affect, how a user determine trust in the content provided by Web information sources [3]. Therefore, the proposed system does filtering of website's trustworthiness based on five major areas as mentioned below [4]:

1. Authority – domain specific
2. Related resources – links from trusted websites
3. Popularity- most visited websites
4. Age - lifespan of time-dependent information
5. Recommendation - referrals from other users

The proposed system, defines eighteen factors which are categorized under the above mentioned five major areas. The trustworthiness of a website is calculated based on these eighteen factors and it is stored thereby increasing the performance in retrieving the trustworthy websites. All the defined eighteen factors are categorized in those major five parameters which can be shown in below Figure1:

The Authority parameter is calculated by determining a high co-occurrence of keywords or semantic phrases from multiple pages of website. This can be done by analyzing the URL. Different weights are assigned depending on the domain names like Page Title, Meta keyword and Meta description value gained for each URL.

The Age parameter consists of two aspects; firstly the last modified date and second aspect is the age of domain. For information's point of view, the newer the content the better. So, to calculate this, a comparison is done between the current date and the last modified date. However, for domain age, it is clear that a long-lasting running website is strongly recommended.

The Popularity parameter is calculated using the number of "In" links, which refer to the number of times a particular website is referenced from other trustworthy websites like Google Page Rank. It would also mean that more the number of visitors to a particular website, more popular that particular website is. Basically the more good quality links a website has, the better it is.

The Related Links parameter is used by adding an appropriate weight to each URL's trustworthiness that is listed out in a Highly-Trustworthy website Like Google inbound links, Yahoo inbound links, Bing inbound links and Alexa inbound links.

The Recommendation is basically when one praises or commend to other as being worthy or desirable. In the same way, when any trustworthy website recommends other website, the value of the recommended website increases. This parameter is calculated based on factors like Alexa Rank, WOT Rating, Site Adviser Rating and Dmoz Listing. These eighteen factors are used to calculate the website trustworthiness.

In the proposed system, the filtering of the trustworthiness is based on all the above mentioned eighteen factors, which would give trustworthy websites as top results.

4. PROPOSED SYSTEM

The proposed architecture for the Trustworthy System is given in Figure 2. The system provides an interface where the user can write his/her query. Once the search button is clicked, it gives the list of URLs on the same page .In meantime; it saves all the URLs in the database.For each URL, all mentioned eighteen factors are calculated and saved as total_score in the database. Then based on the total_score value of each URL, they are rearranged in the descending order, which means if URL has high total_score value then it

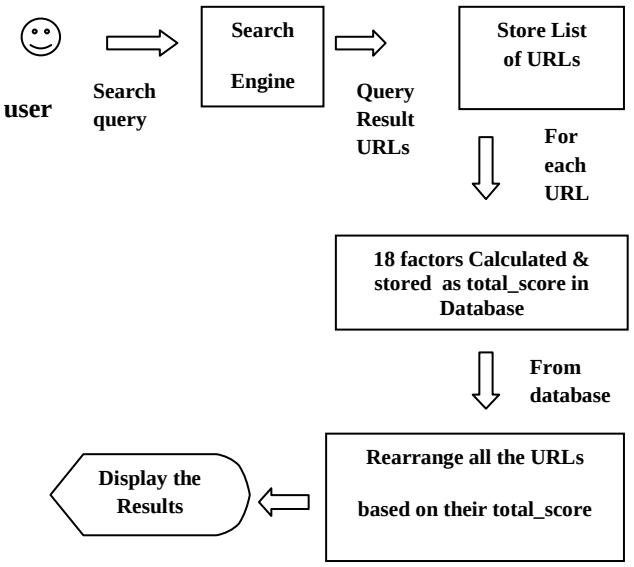

Figure 2: Proposed System Architecture

will appear as the top most result and accordingly rest comes as per their total_score value.

5. ANALYSIS OF EIGHTEEN IMPACT FACTORS

Once user entered his/her query, list of URLs will be available on the same screen. Then each URL is extracted to retrieve the eighteen factors value for calculating them. Then it stores all the calculated factors in the database. Next and final step is that it shows the new result (URLs) in descending order of their Total score. That means if record has high score then it will come to the top and rest comes as per the total score value. Calculations of all 18 factors are as follows:

1. **Page Title:**

   Every html document must have a TITLE Element in the head section. From a search engine's point of view, page title is the first indication of the contents of the page. Additionally, page title is the key information returned when search engines list results to a keyword search [9, 10]. If the searched keyword is in the title tag, then it means the page is most matched.

2. **Meta Keyword:**

   A Meta keywords tag is supposed to be a brief and concise list of the most important themes of the webpage. Long ago in Internet time, the Meta keywords tag was very useful in helping pages to win on search engines. But many unscrupulous webmasters have abused the Meta keywords tag that search engines have had to de-emphasize their importance. Though Meta keywords tags are not a major factor search engines consider when ranking sites, they should not be left off the page [11]. If the searched keyword is in the Meta tag then it means the page is most matched.

3. **Meta Description:**

   The Meta description tag is indented to be a brief and concise summary of web page's content. The Meta description tag is designed to provide a brief description of the website which can be used by search engines or directories. If the searched keyword is in the Meta description then it means the page is most matched [12].

4. **Alexa Rank**:

   Alexa is a very powerful tool used to rank web site traffic. This is one of the most accurate freely available tools to find out how well your site ranks up against millions of other sites on the web. The lower the Alexa ranking number the more heavily visited the site [13].

5. **Google Page Rank**:

   Google Page Rank is one of the methods Google uses to determine relevance or importance of a page. It matters because it is one of the factors that determine a page's ranking in the search results.If the page rank of a particular page is closer to 5 then it is considered as more popular [15].

6. **Google Inbound Links**:

   Defines number of third party websites link to particular website that is identified by Google. Back links are important for SEO because some search engines, especially Google, will give more credit to websites that have a good number of quality back links, and consider those websites more relevant than others in their results pages for a search query. More number of Google inbounds links then more Weight-age is given to it [16].

7. **Google Indexed Pages:**

   Google Indexed Pages gives an indication of number of pages indexed by Google and available in Google servers. There are various on-page SEO factors helping to get higher search engine rankings including the number of web pages indexed by Google and other search engines (indexation). It plays an important role in the SEO score of particular site. In many cases, the winning factor of a site compared to its competitor is the number of its pages indexed by Google or other search engines.[17]

8. **Last Modified Date:**

   Last Modified Date is the date of the particular page created or modified by the website owner. From an informational point of view, newer content is usually better than older content within a web page. When ever user tries to search some information and anticipate that newer information rather than information that is few-years-old [18].

9. **Domain Age:**

   Domain Age defines how old a particular website is. A long-lasting and well-kept domain name on the cyberspace reflects its importance to search engine optimization [18].

10. **Yahoo Inbound Links:**

    Yahoo Inbound Links defines number of third party websites link to particular website that is identified by Yahoo [16]. More number of Yahoo inbound links then more Weight-age is given to it.

11. **Web Of Trust (WOT) Rating:**

    WOT ratings are powered by a global community of millions of trustworthy users who have rated millions of websites based on their experiences. Website's reputation rating is based on ratings from the WOT community, tells

how much other users trust your site. It defines that particular site is safe or not as 'Trustworthy', 'Mostly', 'Suspicious', 'Untrustworthy', 'Dangerous', 'Unknown' and depending upon this Weight-age is assigned to it[19].

12. **Yahoo Indexed Pages:**

Yahoo Indexed Pages gives an indication of number of pages indexed by yahoo and available in yahoo servers [20]. If more pages are indexed by Yahoo and available in Yahoo servers, more it is better.

13. **Alexa Inbound Links:**

Alexa Inbound Links defines number of third party websites link to particular website that is identified by Alexa. Quality inbound links are an essential element of web site marketing and search engine optimization programs to increase traffic and online sales. The greater the number of relevant and authoritative links to a web page, the greater the potential for higher search engine rankings and qualified traffic [13, 14]. More number of Alexa inbounds links then more Weight-age is given to it.

14. **Dmoz Listing:**

Dmoz Listing is the Open Directory Project (ODP) is a multilingual open content directory of WWW links. That means site comes under the particular category. DMOZ is the most respected online directory. All major search engines like Google, Yahoo, and Bing give a lot of importance to websites having links from DMOZ. Inclusion in DMOZ can dramatically increase your website ranking and traffic [21]. If the site comes under the ODP category that means marks will be 100%.

15. **Site Advisor Rating:**

Ratings from SiteAdvisor.com are based on a variety of measures. It claims to protect user by labeling Web sites green, yellow, or red to indicate that they are safe, questionable, or dangerous [22]. Site Advisor Rating defines that site is good (Green), bad (Red), compromised (Yellow) and (Grey). Depending upon the color Weight-age is assigned.

16. **Bing Indexed Pages:**

Bing Indexed Pages gives an indication of number of pages indexed by Bing and available in Bing servers [24].If more pages are indexed by Bing and available in Bing servers, more it is better.

17. **Bing Inbound Links:**

Bing Inbound Links defines number of third party websites link to particular website that is identified by Bing [24].More number of Bing inbounds links then more Weight-age is given to it.

18. **Ask Indexed Pages:**

Ask Indexed Pages gives an indication of number of pages indexed by Ask and available in Ask servers [25]. If more pages are indexed by Ask and available in Ask servers, more it is better.

For each of the mentioned eighteen factors, administrator will provide weight-age based on the importance of each factor. Weight-age would be given to each of the 18 factors which are normalized to 100%. For each saved URL, all 18 factors are calculated and sum it as total score value.

6. CONCLUSION

The WWW is the most important source of information. But, there is no guarantee for information correctness and lots of conflicting information is retrieved by the search engines and the quality of provided information also varies from low quality to high quality.

The proposed system provide trustworthy websites for queries in web searching by filtering of website trustworthiness based on the mentioned eighteen factors and it is stored thereby increasing the performance in retrieving more trustworthy websites. Since proposed system provides more trustworthy websites as Top results, it would save considerable amount of searching time.

7. REFERENCES

[1] Xiaoxin Yin, Jiawei Han, Senior Member, IEEE, and Philip S. Yu, Fellow, IEEE, "Truth Discovery with Multiple Conflicting Information Providers on the Web", Los Angeles, CA, USA, VOL. 20, NO. 6, JUNE 2008, pp. 796-808

[2] Xin Luna Dong, Laure BertiEquille, Divesh Srivastava, "Truth Discovery and Copying Detection in a Dynamic World", VLDB '09, August 2428, 2009.

[3] Sumalatha Ramachandran, Sujaya Paulraj, Sharon Joseph and Vetriselvi Ramaraj, "Enhanced Trustworthy and High-Quality Information Retrieval System for Web Search Engines", IJCSI International Journal of Computer Science Issues, Vol. 5, October 2009, pp38-42.

[4] Gil, Y. and Artz, D.,"Towards content trust of web resources", Edinburgh, Scotland, May 23 - 26, 2006, DOI=http://doi.acm.org/10.1145/1135777.1135861 ,NY, pp565-574.

[5] Soo Young Rieh,"Judgment of Information Quality and Cognitive Authority in the Web", citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.8991

[6] "Page rank algorithm", www-personal.ksu.edu/~eddery/Google.pdf

[7] "Page rank algorithm", http://www.markhorrell.com/seo/pagerank.html

[8] "HITS", http://en.wikipedia.org/wiki/HITS_algorithm#Algorithm

[9] "Page Title", http://www.seologic.com/faq/title-tags

[10] "Significance of page title", http://www.knowthis.com/principles-of-marketing-tutorials/internet-marketing/page-title-for-sem/

[11] "Meta keywords " , http://www.seologic.com/faq/meta-keywords

[12] "Meta description", http://www.seologic.com/faq/meta-descriptions

[13] "Alexa rank", http://developers.evrsoft.com/find-traffic-rank.shtml

[14] "Alexa rank", http://kisswebmaster.com/importance-of-alexa-ranking-why-should-we- increase-it/

[15] "Google Page rank", http://www.webworkshop.net/pagerank.html

[16] "Google inbound links", http://googlewebmastercentral.blogspot.com/2008/10/good-times-with-inbound-links.html

[17] "Google indexed pages", http://www.googleguide.com/google_works.html

[18] "Last modified date and Domain age", http://www.raidenhttpd.com/en/manual/seo.html

[19] "WOT Rating", http://www.mywot.com/

[20] "Yahoo indexed pages", http://www.webmastertools.info/tools/yahoo-indexed-pages-checker/

[21] "Dmoz Listing", http://www.submitedge.com/dmoz_listing.html

[22] "Site adviser rating", http://windowssecrets.com/top-story/siteadvisor-ratings-may-be-1-year-out-of-date/

[23] "Significance of SEO Factors", http://www.dotsandcoms.us/data2/analysis_report.pdf

[24] "Bing indexed pages", http://www.x10tools.com/tools/bing-indexed-pages-checker/

[25] "Ask indexed pages", http://www.ask.com/wiki/Category:Indexed_pages

[26] Haider A. Ramadhan and Khalil Shihab,"A Heuristic Based Approach for Increasing the Page Ranking Relevancy in Hyperlink Oriented Search Engines: Experimental Evaluation",International Journal of Theoretical and Applied Computer Sciences, Volume 1 Number1(2006) pp.49-62