International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 9, September 2016)

Topical and Trust Based Page Ranking

L. Smitha[1], Dr. S. Sameen Fatima[2]
[1]PhD. Scholar, OUCE, Osmania University, Hyderabad, India
[2]Professor, Dept of CSE, OUCE, Osmania University, Hyderabad, India

***Abstract*-- Vast growth and the broad accessibility of data on the web have driven the surge of exploration movement in the zone of data retrieval on the Web. Topic and Trust are the important factors for data retrieval framework. As the extent of the web is huge, it is difficult to fulfil the user's need. To this end, the Web offers a rich site of data, which communicates through the hyperlinks. This paper explains the idea of improving PageRank, "Upgraded Page Rank with Topic utilizing Trust Component" which has the vast limit as compared with Conventional Page Rank Algorithm.**

***Keywords*-- Ranking, Page Rank, Topic sensitive PageRank, Backlinks, Hyperlinks, and Trust ranking.**

I. INTRODUCTION

Searching on the World Wide Web is the most frequent operation on the Web. Therefore, it is essential to have search engines for finding required information on the web. Recently, vast research is going on in the field of information technology to producing good search engines that can search efficiently and effectively. Researchers are publishing rich literature in proximity of information retrieval [2], the size of the web increases every day and users on the web are diverse who perform search, postures new challenges.

II. RELATED WORK

In this paper, we combine a notion of Topic-specification with Trust using Dirichlet PageRank and Trust-based Ranking Algorithm [14]. This idea taken from the PageRank paper by Larry Page and Sergey Brin (1998), who debates about the personalization of PageRank by presenting a bias towards only some trusted web sites. Some content providers can easily make high quality pages to increase the ranking in the web. Some try to change the features of pages; this is Web spam [24, 12]. Henzinger et al. [3], declares that search engine spam is one of the major challenges in search engine development. Several researchers followed Brin, Page and Haveliwala in changing only the bias probabilities, including Wu et al. (2006). Many researchers have studied the types of web spam, and their work can be summered here. D. Fetterly, et al. [9]. A. A. Benczur, et al. - propose SpamRank [13] fully automatic link spam detection in which for every page, the calculation of PageRank is to be done for all incoming links. Wu and Davison studied the combination of both incoming, outgoing links and a broadcast step to notice link farms.

Drost and Scheffer [8] gave a machine learning method to discover link spam. From the beginning, the knowledge of a focused or custom PageRank vector has occurred [8], Haveliwala [10] initially intended to get topical information into PageRank calculation. In this paper, different topics are required to be selected and then apply PageRank to find good pages within topical networks.

III. PAGE RANK

PageRank [12] is an eminent algorithm that is based on considering the link information for giving all the pages global significance scores. Larry Page originally gave this proposal [4] in which PageRank of a page becomes important if the number of other important web pages is pointing to that page. Consistently, this will be based on a mutual support between pages. Here, the importance of page effects and is effected by the rank of other pages in the web. The PageRank score **r** (*p*) of a page *p* is given as:

$$\mathbf{r}(p) = \alpha \cdot \sum_{q:(q,p)\in\varepsilon} \frac{\mathbf{r}(q)}{\omega(q)} + (1-\alpha)\cdot\frac{1}{N}, \quad \text{----- 1}$$

Later, we can sum up two elements to get the importance of a page p: one element is from the score comes from incoming links to p and the other (static) element is equal for all web pages.

$$r = \propto T.\mathrm{r} + (1-\propto).\frac{1}{\mathrm{N}}.1_N \quad \text{----- 2}$$

A key factor to be considered is, the regular PageRank algorithm allocates an alike static score to each page, this rule is cancelled in a biased PageRank version. PageRank in the form of matrix equation,

$$r = \propto T.\mathrm{r} + (1-\propto).\mathrm{d} \quad \text{----- 3}$$

d is a static score distribution vector of random, nonnegative entries adding to get total as one.

PageRank is the famous algorithm implemented by Google to help define where websites should rank in their search engine index. PageRank can measure the status of a page. PageRank works by calculating the number of inbound hyperlinks to a particular page. The primary hypothesis is that, websites that are more reliable are likely to get more links from other websites. If we consider each link as a vote from another web master, Google applies this data to determine if a site has good contents and promotes it accordingly.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 9, September 2016)

IV. TOPIC SENSITIVE PAGERANK

In Topic-Sensitive PageRank approach, shown in the figure 1, importance scores like PageRank offline are to be calculated. Like this, multiple importance scores are computed for every page then the calculation of such set of scores for finding the importance of the page is done considering many domains. At query time, depending upon the topic of the query, adding of the importance scores together. This will result as a composite PageRank score for the pages similar to the query. The same score can be used in aggregation with other similarity based schemes to yield a final rank for the result pages depending on the given query.

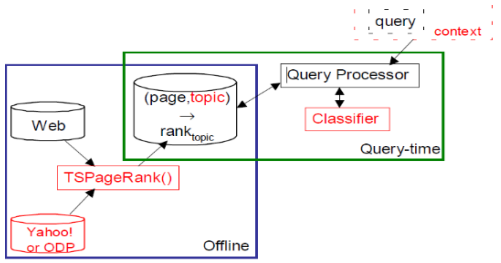

Fig 1: Topic Sensitive Page Ranking

In the primary stage, a set of biased vector is generated using the set of topics. This stage is completed once, offline, at the time of preprocessing of the crawling. For the described URL, personalization vector $\vec{p}$ can be used in the various categories in the ODP, 16 different biased PageRank vectors are created. In certain condition,

The set of URLs in the ODP category be: T_j .

Different topics or categories present in ODP are: c_j

Then when calculating the PageRank vector for topic c_j, that replaces the uniform damping vector, we use the nonuniform vector

Where $\vec{p} = [\frac{1}{N}]_{N\times 1}$

$$M' = (1-\alpha)M + \alpha[\frac{1}{N}]_{N\times N} \quad \text{---- 4}$$

$$\vec{Rank} = M' \times \vec{Rank} \quad \text{----5}$$

$$= (1-\alpha)M \times \vec{Rank} + \alpha\vec{p} \quad \text{----6}$$

The PageRank vector for topic c_j will be denoted as, the single unbiased PageRank vectors are generated (indicated as NoBias) for comparison. There is computation done for 16 class term-vectors having terms in the documents from each of the 16 top-level categories. D_{jt} gives the total summation of occurrences of term t in documents listed below class c_j of the Open Directory Project. Other sources are used to visualize and for creating topic sensitive PageRank vectors. Open Directory Project is compiled by thousands of volunteer editors, is less inclined to any topic or cannot be biased.

During the time of Query, the second step is performed, if a given query q, let q_0 be the context of q. Class probabilities are computed for each of the 16 top-level ODP Topics, trained on q_0. We assume q'_i is the i_{th} term occurring in the query or context q′. Then specified q is the query, for every c_j, the subsequent computed is shown:

$$P(c_j|q') = \frac{P(c_j)\cdot P(q'|c_j)}{P(q')} \propto P(c_j)\cdot \prod_i P(q'_i|c_j) \quad \text{---- 7}$$

Lastly, computing the query sensitive importance scores for each of these rescued URLs as follows: The $rank_{jd}$ be the rank of the document given by the rank vector for the document d, the query-sensitive importance score value $score_{qd}$ as follows.

$$\sum_j [w_j \vec{PR}(\alpha, \vec{v_j})] = \vec{PR}(\alpha, \sum_j [w_j \vec{v_j}]) \quad \text{----- 8}$$

The results are ranked according to this composite score value $score_{qd}$. The PageRank computation has the following probabilistic interpretation, with respect to random surfer model [7]. Consider w_j as the coefficient used to weight the j_{th} rank vector, with $\sum_j w_j = 1$. Let $w_j = (c_j|q)$, Then note that the equality holds. Thus topic-sensitive score value $score_{qd}$ is generated as the random walk on the Web. With probability of $1-\propto$, a random surfer on page u follows outlink of u (where the particular outlink is chosen randomly). With probability $\alpha P(c_j|q`)$ to one of the pages in Tj, the surfer instead jumps (where the particular page in T_j is chosen uniformly at random). Once webmasters figured out that links were the key to higher rankings, they began business with the links and started giving high ranks in Google's search results process and thus TrustRank was born.

Google continues to attempt to deliver more relevant search results in the year 2014 Matt Cutt referred to an improved version of PageRank (Topical PageRank). Google gives user matched documents for a query, but what factors it considers is not known to achieve this. Google is trying to achieve this and require too much CPU run time to execute. The other approach that Google is trying to attempt is to classify websites by topical theme into different categories. Information on Topical PageRank is still rather under research.

Unlike regular PageRank, "Topical PageRank" is a measure of "Authority." Whereas regular PageRank can be used to rank a website in the more competitive position. "Topical PageRank" can make the difference between the position of first page and allocating rank to a site at the top of the first page. In fact, the research is done on position, we see topical patterns appearing and identification of topical classification can be done, this is featured in Google's top SERPS.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 9, September 2016)

V. Adjusting Rank Based On Trust

While it is interesting to be able to devise ranking systems that take known spammers into account, it is also important to calculate a ranking based on various concepts of trust in a network. There are numerous scenarios to consider, and Dirichlet PageRank with boundary conditions .[15].will be a useful algorithmic tool.

$$\text{Score}_{qd} = \sum_{c} P(c_j|q') \cdot rank_{jd}$$

Consider the a situation in a network G, node v wants to compute a personalized ranking of the nodes, but v trusts its own friends and wants its ranking on the top þ fraction of nodes to be similar to its friends'. Presumably decisions. Vertex v can efficiently compute a personalized PageRank vector as its ranking function using algorithms from [14], but PageRank alone will not take into account the implied trust between v and its friends. But using Dirichlet PageRank with boundary conditions, we can take v's trusted friends into account. We illustrate this in the algorithm 2[14].

Selecting the seed set becomes basic part of the functioning of trust propagation based algorithms, e.g., Trust Rank [5]. While selecting the seed set physically or by humans, then the seed set becomes limited to a small size, as it is almost impractical to create a large size seed set manually. The seed set of small-sized can badly affect the final search engine ranking result pages (SERPs). Therefore, we can expand the initially manually selected seed set to a large one by using automation mechanism. The Automatic seed set expansion (ASE) [8] using joint the recommendation link structure seed set automatic expansion takes place.

Conventionally the human expertise evaluation process of web pages, have limitations both in terms of quality and quantity. The performance of anti-spamming algorithms is affected by the quality and quantity of seed sets, which causes serious effects. If the initial seed set is not chosen properly, then miscalculation leads in propagation. The size of the seed set is also a vital in choosing the seed set. If the seed is small, then top-ranked results will be filled with seeds. A small seed set is inadequately descriptive to wrap diverse topics on the web, and it would create topic bias. Because there are so many reasons we need to expand the seed set by choosing imminent qualified seeds. The (ASE) Automatic seed expansion algorithm utilizes a combined recommendation link structure to choose seeds for expansion.

A. The Impact of the Seed Set

There are drawbacks with a small seed set and manually selected seed set.

Initially, the result of topical and trust ranks are strongly biased towards the high-ranked pages involved in the initial seed set. In spite of propagating trust to other trusts (due to high quality pages), the effect is attenuated because of trust damping [1]. The next reason is if the seed set contains pages concern to some topic or field; it would thus the source topic bias towards these topics. Pages with the bulky portion domain have highest scores. Lastly, the performance of trust rank also declines because of some possible links from reputable to spam pages.

VI. Conclusion

In this paper, an upgraded Page rank calculation utilizing trust using Dirichlet PageRank and Trust-based Ranking Algorithm has been proposed. A Relative investigation of the computational aspects of the proposed plan with the past work means that the proposed Topical Trust PageRank is a superior option to the formerly presented calculation. In this paper, experiments are presented that pinpoint the modifications to PageRank are necessary to adequately cater for the highly specialized situation we encounter in science and technology. This results in a general ranking model for technology incorporating a robust concept of authority. This implementation needs only slight resources and depends only on trust and PageRank calculation, which means that it is efficient during training and at the search time.

In the paper, two evaluations are performed; one is Nobiased and another topical shown in figure 2 and 3. In both the model considerably outperforms not the only state of-the-art, but also standard PageRank, PageRank, and non-topical PageRank. This model achieves its good performance by using only the raw text and links. We wanted an entirely independent authority-based IR model similarity. If the reader wishes to evaluate the performance of Topical and trust based ranking on some PDF papers, it has been incorporated into the software called Qiqqa reference management.

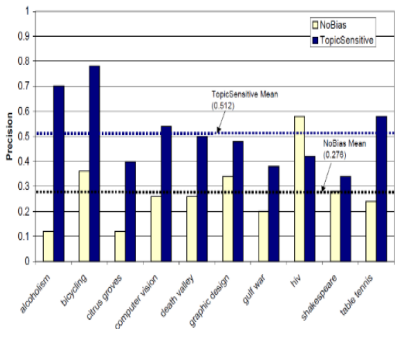

Fig 2: Nobiased and topical results

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 6, Issue 9, September 2016)

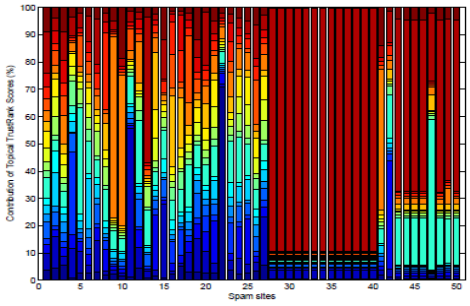

Fig 3: Topical contribution in Algorithm for spam sites

In this paper, two evaluations are performed: one is Nobiased and another topical. In both the model considerably outperforms not the only state of-the-art, but also standard PageRank, PageRank, and non-topical PageRank. This model achieves its good performance by using only the raw text and links. We wanted an entirely independent authority-based IR model similarity. If the reader wishes to evaluate the performance of Topical and trust based ranking on some PDF papers, it has been incorporated into the software called Qiqqa reference management.

VII. Future Work

During the time of Query, for a given query, Class probabilities are computed for each of the 16 top-level ODP Topics and trained but as a finer grained topics can be used and next level of hierarchies can be considered for calculating vectors in ranking .This may work as future enhancement for finer levels of topics in the topical and trust ranking Algorithm.

REFERENCES

[1] Baoning Wu and Brian D. Davison (2005), Identifying Link Farm Spam Pages, Proceedings of the 14th International World Wide Web Conference, Chiba, Japan, pp. 820-829.

[2] Neelam Duhan, A.K.Sharma and Komal Kumar Bhatia,"Page Ranking Algorithms: A Survey", in proceeding of the IEEE International Advanced Computing Conference (IACC), 2009.

[3] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. SIGIR Forum, 36(2):11{22, Fall 2002.

[4] S.Brin and L.Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, Vol 30, Issue 1-7, 1998.

[5] Baoning Wu and Brian D. Davison (2005), Identifying Link Farm Spam Pages, Proceedings of the 14th International World Wide Web Conference,Chiba, Japan.

[6] C. Ridings and M. Shishigin, "Pagerank Uncovered", Technical] Report, 2002.

[7] Tamanna Bhatia, "Link Analysis Algorithm for Web Mining", IJCST Vol 2, Issue 2, June 2011.

[8] Xianchao Zhang, Bo Han, Wenxin Liang (2013), Automatic Seed Set.

[9] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In Proceedings of WebDB, pages 1-6, June 2004.

[10] T. Haveliwala. Topic-sensitive PageRank. In Proceedings of the Eleventh International World Wide Web Conference, pages 517{526, Honolulu, Hawaii, May 2002.

[11] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In Proceedings of the 13th International World Wide Web Conference, pages 403{412, New York City, May 2004.

[12] Sergey Brin and Lawrence Page (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine., Computer networks and ISDN Systems, pp. 107-117.

[13] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher. SpamRank - fully automatic link spam detection. In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005.

[14] F. Chung and W. Zhao. A sharp PageRank algorithm with applications to edge ranking and graph sparsifcation. Proceedings of Workshop on Algorithms and Models for the Web Graph (WAW 2010), Lecture Notes in Computer Science 6516,2-14.

[15] Dirichlet PageRank and Trust-based Ranking Algorithms Fan Chung, Alexander Tsiatas, and Wensong Xu Department of Computer Science and Engineering University of California.