iJRASET
International Journal For Research in
Applied Science and Engineering Technology

Volume: 6 Issue: I Month of publication: January 2018

DOI: http://doi.org/10.22214/ijraset.2018.1186

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor :6.887
Volume 6 Issue I, January 2018- Available at www.ijraset.com

An Effective Link Algorithm for Web Mining Based on Topic Sensitive

R. Suganya
Assistant Professor, Head Department of Computer Science, Annai Vailankanni Arts and Science College, Thanjavur, Tamilnadu, India

Abstract: With the growing on internet technology, each service provider to provide proper, relevant and quality of information to the internet users against their query submitted to the search engine. Web Structure Mining (WSM) and Web Content Mining(WCM) plays an important role in this approach. Some page ranking algorithms Page rank, Weighted Page Content Rank(WPCR), Topic Sensitive Page Rank(TSP) are commonly used for WSM and WCM. To get more accurate result a new algorithm proposed for rank the results of search page based on users topic or query. We propose Topic Sensitive Weighted Page Content Rank(TSWPCR) algorithm based on Web structure mining and content mining, this will provide better result compared with other page ranking algorithms. For ordinary search queries, TSWPCR will satisfy the topic sensitive of the query.

Keywords: WSM, WCM, WPCR, TSP, TSWPCR

I. INTRODUCTION

The volume of information available on World Wide Web[4][2], it expanded in size and complexity. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. Huge amount of information becomes very difficult for the users to find, extract, filter or evaluate the relevant information.

Web mining can be easily executed with the help of other areas like Database(DB), Information Retrieval(IR), Natural Language Processing (NLP), and Machine Learning etc.

The World Wide Web serves as a huge , widely distributed , global information service for news, advertisements, consumer information, financial management, education, government, e-commerce and many other information services.

This paper is organized as follows : Section II Web Mining Process defined, in Section III Web Mining Categories are explained, Section IV describes the search engine architecture , Section V(A) defines the Weighted Page Content Rank Algorithm, Section V(B) defines the Topic Sensitive Pagerank Algorithm. Section VI describes the modified search engine architecture and Section VII defines the proposed algorithm TSWPCR , Section VIII comparative study of proposed and previous algorithms and Section IX concluded.

II. PROCESS OF WEB MINING

The complete process of extracting knowledge from Web data[1] is as follows in Fig.1.

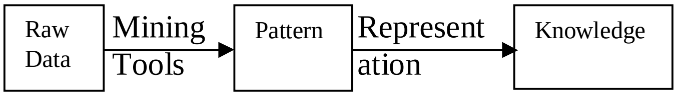


Fig. 1: Web Mining Process

According to Kosala et al[3]Web mining can be decomposed into the subtasks, namely:

1) *Resource finding*: It is the task of retrieving intended web documents.
2) *Information selection and pre-processing*: Automatically selecting and pre- processing specific from information retrieved Web resources.
3) *Generalization*: Automatically discovers general patterns at individual Web site as well as multiple sites.
4) *Analysis*: Validation and interpretation of the mined patterns.

III. WEB MINING CATEGORIES

Web mining aims at finding and extracting relevant information that is hidden in Web-related data, in particular hypertext documents published on the Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Usage Mining, and Web Structure Mining as shown in Fig.2.

A. Web Content Mining

Web Content Mining [3] [8] [6] is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. Web content mining is related but is different from data mining and text mining. Web content mining is also different from text mining because of the semi-structure nature of the web, while text mining focuses on unstructured texts. The technologies that are normally used in web content mining are NLP (Natural
language processing) and IR (Information retrieval).

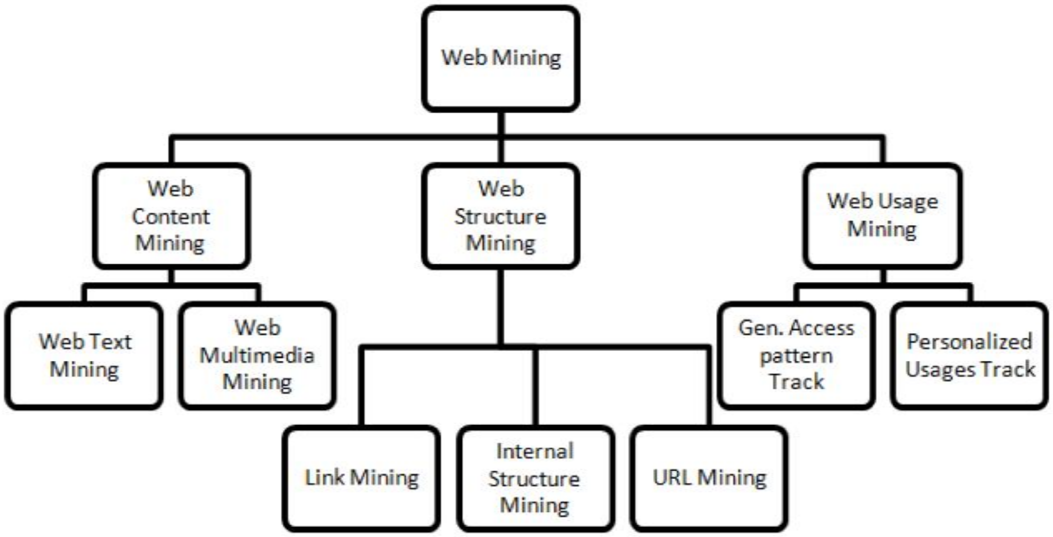


Fig.2. Web Mining Categories

B. Web usage mining

Web usage mining [3][5] is the application of data mining techniques to discover usage patterns from Web data in order to understand and better serve needs of Web based applications. It consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. However, one of the major challenges faced by Web usage mining applications is that Web server log data are anonymous, making it difficult to identify users and user sessions from the data.

C. Web structure mining

The goal of web structure mining [7] is to generate structural summary about the website and web page. The first kind of web structure mining is extracting patterns from hyperlinks in the web. A hyperlink is a structural component that connects the web page to a different location. The other kind of the web structure mining is mining the document structure. It is using the tree-like structure to analyze and describe the HTML (Hyper Text Markup Language) or XML (Extensible Markup Language).

IV. SEARCH ENGINE ARCHITECTURE

Search engines [10] are the key to finding specific information on the vast expanse of the World Wide Web. We use the term search engine in relation to the Web. These usually refer to the actual search forms, which searches through databases of the HTML documents. The Search Engine Architecture shown in Fig.3. Here we represent the elements of search engine architecture.

1) *User Search:* The users do besides typing a few relevant words into the search form. Can they specify those words which must be in the title of a page? About specifying those words which must be in an URL. Most engines allow you to type in a few words, and then search for occurrences of these words in their data base.
2) *Crawlers:* A crawler is a program that visits Web sites and reads their pages and other information in order to create entries for a search engine index. The search engines on the Web have a program, which is also known as a "spider" or a "bot." Crawlers are typically programmed to visit sites that have been submitted by their owners as new or updated. Entire sites or specific pages can be selectively visited and indexed. Crawlers apparently gained the name because they crawl through a site a page at a time, the links to other pages on the site until all pages have been read.

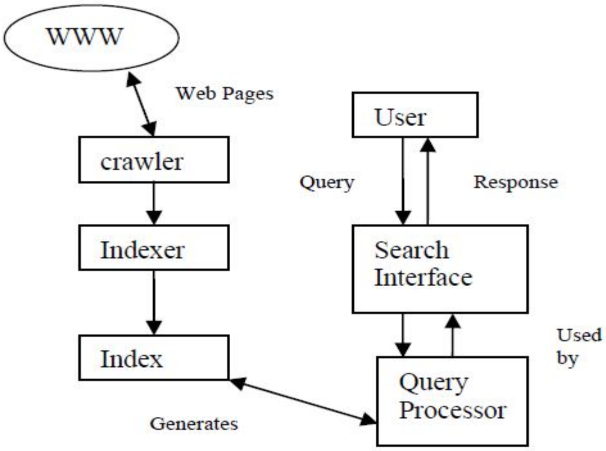


Fig. 3 Search Engine Architecture

3) *WWW:* The World Wide Web is: all the resources and users on the Internet that are using the Hypertext Transfer Protocol. The World Wide Web is a system of interlinked hypertext documents accessed via the Internet. Web that may contain text, images, video, and other multimedia and navigates between them using hyperlinks.
4) *Indexer:* (“Internet indexing") provide a more useful vocabulary for Internet or onsite search engine. With the increase in the number of periodicals that have articles online, web indexing is also becoming important for periodical websites.

V. PAGE RANKING ALGORITHMS

World Wide Web is large sized repository of interlinked hypertext documents accessed via the Internet. The user navigates through this using hyperlink. Search Engine gives millions of results and applies Web mining techniques to order the results. The sorted order of search results is obtained by applying some special algorithms called—Page ranking algorithms. There are two Page Ranking algorithms; Weighted Page Content Rank and Topic Sensitive PageRank. They are the commonly used algorithm in Web Structure Mining and Web Content Mining.

A. Weighted Page Content Rank

Weighted Page Content Rank Algorithm (WPCR) is a page ranking algorithm which is used to give a sorted order to the web pages returned by a search engine in response to a user query. WPCR is a numerical value based on which the web pages are given an order. This algorithm employs web structure mining as well as web content mining techniques. Web structure mining is used to calculate the importance of the page and web content mining is used to find how much relevant a page is? Importance here means the popularity of the page i.e. how many pages are pointing to or are referred by this particular page. It can be calculated based on the number of inlinks and outlinks of the page. Relevancy means matching of the page with the fired query. If a page is maximally matched to the query, that becomes more relevant.

The formula to calculate the Weighted Page Content Rank of a page U is given in Eq. 1.

$$PR(U) = (1-d) + d\sum_{v\in B(u)} PR(V)W^{in}(U,V)W^{out}(U,V) * \quad (C_w + P_w) \qquad (1$$

Where,

PR(U) = Pagerank of page U
B(U) = Set of all pages referring to page U
D = Damping factor which can be set between 0 and 1,
W^{in}(U,V) = inweight of link (U,V)
W^{out}(U,V) = outweight of link(U,V),
C_w = Content weight of page U
P_w = Probability weight of page U

B. Topic-sensitive pagerank

In topic-sensitive Page Rank precompute the importance scores offline, as with ordinary PageRank. However, we compute multiple importance scores for each page; we compute a set of scores of the importance of a page with respect to various topics. At query time, these importance scores are combined based on the topics of the query to form a composite PageRank score for those pages matching the query. As the scoring functions of commercial search engines are not known, in our work we do not consider the effect of these IR scores (other than requiring that the query terms appear in the page). We believe that the improvements to PageRank's precision will translate into improvements in overall search rankings, even after other IR-based scores are factored in.

1) *ODP-biasing:* The first step in is to generate a set of biased PageRank vectors using a set of basis topics. This step is performed once, offline, during the reprocessing of the Web crawl. There are many possible sources for the basis set of topics. However, using a small basis set is important for keeping the reprocessing and query-time costs low. One option is to cluster the Web page repository into a small number of clusters in the hopes of achieving a representative basis. We chose instead to use the freely available, hand constructed Open Directory as a source of topics.

Let T_j be the set of URLs in the ODP category $c_{j.}$ Then when computing the PageRank vector for topic c_j , in place of the uniform damping vector $p = \left[\frac{1}{n}\right] n \times 1$, we use the nonuniform vector p = v_j where

$$v_{ji} = \begin{cases} \frac{1}{T_j} & i \notin T_j \\ 0 & i \notin T_j \end{cases} \qquad (2)$$

The PageRank vector for topic c_j is given by PR(α,v_j). We also generate the single unbiased PageRank vector (denoted as NOBIAS) for the purpose of comparison.

2) *Query-Time Importance Score:* The second step is performed at query time. Given a query q, let q' be the context of q. In other words, if the query was issued by highlighting the term q in some Web page u, then q' consists of the terms in u. Alternatively, we could use only those terms in u nearby the highlighted term, as often times a single Web page may discuss a variety of topics. For ordinary queries not done in context, let q' = q. Using a multinomial naive-Bayes classifier with parameters set to their maximum-likelihood estimates, Then given the query q, we compute for each c_j the following:

$$P(c_j|q') = \frac{P(c_j).P(q'|c_j)}{P(q')} \alpha P(c_j).\prod P(q'_i|c_j) \qquad (3)$$

$P(q'|c_j)$ is easily computed from the class term-vector D_j .The quantity P(c_j) is not as straightforward. We chose to make it uniform, although we could personalize the query results for different users by varying this distribution. In other words, for some user k, we can use a prior distribution $P_k(c_j)$ that reflects the interests of user k.

Using a text index, we retrieve URLs for all documents containing the original query terms q. Finally, we compute the query-sensitive importance score of each of these retrieved URLs as follows. Let r_{jd} be the rank of document d given by the rank vector PR(α,v_j) (i.e., the rank vector for topic c_j). For the Web document d, we compute the query-sensitive importance score s_{qd} as follows in Eq. 4.

$$S_{qd} = \sum_j P(c_j|q').r_{jd} \qquad (4)$$

The results are ranked according to this composite score $s_{qd.}$

The above query-sensitive PageRank computation has the following probabilistic interpretation, in terms of the “\random surfer” model. Let w_j be the coefficient used to weight

the jth rank vector, with$\sum_j w_j = 1$ (e.g., let $w_j = P(c_j|q)$).

Then note that the equality

$$\sum_j[w_j PR(\alpha, v_j)] = PR(\alpha, \sum_j[w_j v_j]) \qquad (5)$$

Thus we see that the following random walk on the Web yields the topic-sensitive score s_{qd}. With probability 1- α, a random surfer on page u follows an outlink of u (where the particular outlink is chosen uniformly at random). With probability P(c_j |q'), the surfer instead jumps to one of the pages in T_j (where the particular page in T_j is chosen uniformly at random). The long term visit probability that the surfer is at page v is exactly given by the composite score s_{qd} defined above. Thus, topics exert influence over the final score in proportion to their affinity with the query (or query context).

VI. MODIFIED SEARCH ENGINE ARCHITECTURE

On the vast expanse of the World Wide Web search engines are the key to find the specific information. There are at least three elements which contain important: information for a search engine: discovery of the database, the user search, presentation and ranking of results. With the proposed TSWPCR, the search engine architecture is modified so as to add the components for calculating importance and relevancy of pages. The modified architecture is displayed in Figure 4.

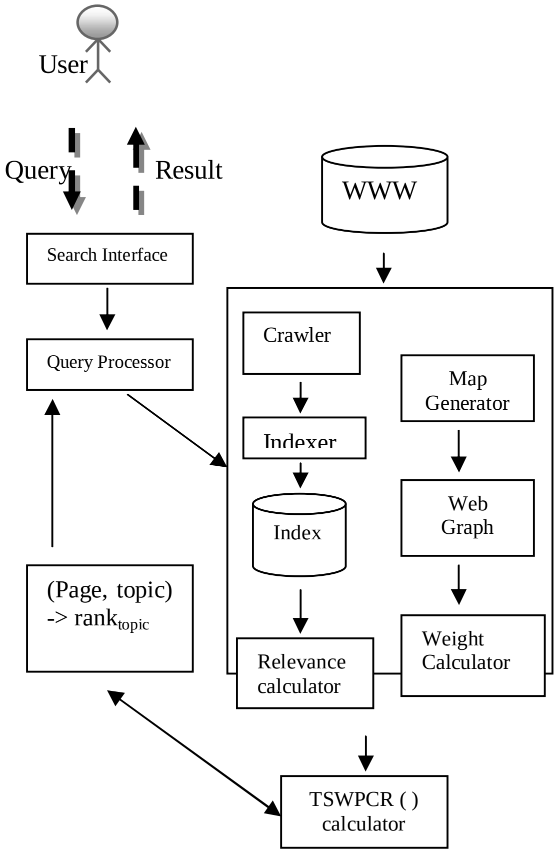


Fig .4. Modified Search Engine Architecture

The Various components and search process are explained below to have an understanding of the existing as well as modified architecture.

1) *WWW:* The World Wide Web is a system of interlinked hypertext documents accessed via the Internet. Web may contain text, images, video, and other multimedia data and user navigates between them using hyperlinks.
2) *Crawler:* A crawler is a program that visits Web and reads their pages and other information in order create entries for a search engine index. The search engines on the Web have crawlers embedded in them. These are typically programmed to visit sites that have been submitted by their owners as new or updated. Entire sites or specific pages can be selectively visited and indexed.
3) *Search Interface*: It is the Graphical interface of a search engine on which the user can enter his query e.g. the Google interface.
4) *Query Processor*: It is the component used for taking the user query from the search interface and processing it word by word.
5) *Indexer:* It provides a more useful vocabulary for Internet or onsite search engine. This component uses HTML parser to extract the terms from web pages after ignoring stop words, prepositions and word stems. The *index* is usually built in an alphabetical order of terms and contains extra information regarding the page such as its URL, frequency and position of terms etc.
6) *Map Generator:* This module generates a map/graphical structure of the WWW. This map is used to further find out the inlinks and outlinks of the web pages.
7) *Weight Calculator:* Weight calculator calculates weight of inlinks and outlinks of the pages. Weight determines how much important a page is on the web.
8) *Relevance Calculator:* Relevance calculator determines the relevancy of the pages to the given query. The page should be matched maximum to the content of the query fired by the user.

9) *TSWPCR Calculator:* It combines the output of weight calculator and relevance calculator to determine the weighted Page Content Rank of all the pages returned after matching the query with the index.

VII. TOPIC SENSITIVE WPCR ALGORITHM

As stated earlier, TSWPCR is a numerical value to represent the rank of a web page. The algorithm to find TSWPCR of a web page is given in Figure 5.

```
Algorithm: Topic Sensitive  WPCR calculator
                    Input:
    T → Topic
    P → Page
    Q → Query
    all backlinks of P → Inlink and Outlink Weights
    $t_1,t_2,\ldots.t_n$ → Topic parameter
    d (damping factor)
Output: Rank score
Step 1: Relevance calculation:
    a) Find T Topic Sensitive of Q
    b) $t_i$ parameters of topic sensitive (where i=1,2,3…..n)
    c) Find all meaningful word strings of Q.
    d) Find whether N strings / ($t_i$) parameters are occurring in P or Not?
    e) Z= Sum of frequencies of all  N + $t_i$ (where i=1,2,…n)
    f) S= Set of the maximum possible strings occurring in P.
    g) X= Sum of frequencies of strings in S.
    h) Content Weight(CW)= X/Z
    i) C= No. of query terms in P
    j) D= No. of all query terms of Q while ignoring stop words.
    k) Probability Weight(PW)= C/D
Step 2: Rank calculation:
    a) Find all backlinks of P(say set B)
    b) $PR(U) = (1-d) + d\sum_{v\in B(u)} PR(V)W^{in}(U,V)W^{out}(U,V) * (C_w + P_w)$
    c) Output PR(P) i.e. the Rank score
```

Fig. 5 TopicSensitiveWPCR Algorithm

The various steps of the algorithm are explained below in detail.

A. Weight calculation

The W^{in}(v,u) and W^{out}(v,u) are the preprocessed weights. Both are just inputted to the algorithm. W^{in}(v,u) is the weight of link(v, u) calculated based on the number of inlinks of page *u* and the number of inlinks of all reference pages of page *v* and is given in Eq (6).

$$W^{in}(U,V) = \frac{I_u}{\sum_{P\in R(V)} I_p} \qquad (6)$$

Where

I_u = number of inlinks of page u

I_p = number of inlinks of page p,

R(v)=Reference page text of page v
The $W^{out}(v,u)$ is the weight of link(v, u) calculated based on the number of outlinks of page *u* and the number of outlinks of all reference pages of page *v* given in Equ (7).

$$W^{out}(U,V) = \frac{O_u}{\sum_{P\in R(V)} O_p} \qquad (7)$$

Where
O_u=number of outlink of page u,
O_p= number of outlink of page p

B. Relevance Calculation

Relevance calculator calculates the relevance of a page on the fly in terms of two factors: one represents the probability of the query in the page and other gives the maximum matching of the query to the page.

Probability Weight: It is the probability of the query terms in the web page. This factor is the ratio of the query terms present in the document and the total number of terms in the
fired query (after ignoring stop words etc.).The formula is given in Eq (8).

Probability weight (P_{wi}) =Y_i/N (8)

where
Y_i= Number of query terms in ith document
N=Total number of terms in query

1) *Content Weight*: It is the weight of content of the web page with respect to query terms. This factor is the ratio of the sum of frequencies of highest possible query strings in order and sum of frequencies of all query strings in order. The maximum possible strings are selected in such a way that all such strings represent a different logical combination of words. The formula for Content Weight is given in Equ (9).

Content weight (C_{wi}) = X_i/M (9)

Where
X_i= Sum of cardinalities of highest possible query strings in order
M= Sum of cardinalities of all possible meaningful query strings in order.

VIII. COMPARATIVE STUDY

Comparison between Page Rank/ Weighted page rank, Weighted Page Content Rank, Topic Sensitive PageRank and Proposed Topic Sensitive Weighted Page Content Rank algorithms show in Fig. 6.

S.No	Page Rank/Weighted Page Rank	Weighted Page Content Rank	Topic Sensitive page rank	Topic Sensitive Weighted Page Content Rank
1	WSM	WSM & WCM	WSM	WSM & WCM
2	They rely on links only	They rely on links and contents	They rely on links	They rely on links and contents based on Topic sensitive
3	Weight of the web page are calculated on the basis of inlinks and outlinks	It gives different weight to web links based on 3 attributes: relative position in page, tag where link is contained , length of anchor text	It computes the rank of page according to the importance of content available on the particular web page	It computes the rank of page according to importance of content available, relative position, length of anchor text and tag where link is contained
4	Minimum determination of the relevancy of the pages to the given query as mentioned	Minimum determination of the relevancy of the pages to the given query as mentioned	Minimum determination of the relevancy of the pages to the given query as mentioned	Maximum determination of the relevancy of the pages to the given query as mentioned
5	Provide less relevant of information	WPCR algorithm provide important information and relevancy about a given query	It provide relevant information to the given query	TSWPCR algorithm provide important information and relevancy about a given query with recommended and filtered topic content.

Fig. 6 Comparison between PR/WPR, TSPR, WPCR, and TSWPCR

IX. CONCLUSION

The popularity of World Wide Web has received a tremendous attention by majority of the people to find and retrieve relevant information for various purposes. Therefore, most of the researchers pay attention to web structure mining and web content mining for extracting relevant document for the query given by users. The proposed algorithm TSWPCR algorithm will provide the effective search result compare than other Page ranking algorithms. This algorithm is very helpful to improve the order of the pages with topic based result list, so that the user gets the important and relevant pages easily in the list.

REFERENCES

[1] Cooley, R., Mobasher, B., and Srivastava, J. "Web mining: Information and pattern discovery on the World Wide Web". In proceedings of the 9th IEEE International conference on Tools with Artificial Intelligence (ICTAI' 97), Newposrt Beach, CA, 1997.

[2] A. A. Barfourosh, H.R. Motahary Nezhad, M. L. Anderson, D. Perlis, Information Retrieval on the World Wide Web and Active Logic: A Survey and Problem Definition, 2002.

[3] R. Kosala and H. Blockeel." Web mining research": A survey. ACM SIGKDD Explorations, 2(1):1–15, 2000.

[4] O. Etzioni. The World Wide Web: Quagmire or gold mine. "Communications of the ACM" 39(11):65-68, 1996.

[5] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan, and Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, ACM SIGKDD Explorations Newsletter, January 2000, Volume 1 Issue.

[6] Wang Jicheng, Huang Yuan, Wu Gangshan, Zhang Fuyan. Web mining: knowledge discovery on the Web. Systems, Man, and Cybernetics, 1999.IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference.

[7] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A.Tomkins, D. Gibson, and J. Kleinberg." Mining the Web's link structure". Computer, 32(8):60–67, 1999.

[8] Raymond Kosala, Hendrik Blockeel, "Web Mining Research": A Survey, ACM SIGKDD Explorations News;letter, June 2000, Volume 2 Issue 1.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: ringing order to the web. Technical report, Stanford Digital Libraries SIDL-WP- 1999-0120, 1999.

[10] C. Ridings and M. Shishigin. Pagerank uncovered. Technical report, 2002.

[11] Taher H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No4, July/August 2003, 784-796.

[12] J. Wang, Z. Chen, L. Tao, W. Ma, and W. Liu. Ranking user's relevance to a topic through link analysis on web logs. WIDM, pages 49–54, 2002. Cooper, C. 8

[13] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1–7):107–117, 1998.

[14] W. Xing and Ali Ghorbani, "Weighted PageRank Algorithm", Proc. ofthe Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[15] N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms:A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009.

[16] M. G. da Gomes Jr. and Z.Gong, "Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference onInformation Acquisition, 2005.

[17] A. Broder, R. Kumar, F Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, "Graph Structure in the Web", Computer Networks: The International Journal of Computer and telecommunications Networking, Vol. 33, Issue 1-6, pp 309-320, 2000.

[18] X. Wang, T. Tao, J. T. Sun, A. Shakery and C. Zhai, "DirichletRank: Solving the Zero-One Gap Problem of PageRank". ACM Transaction on Information Systems, Vol. 26, Issue 2, 2008.

[19] Z. Gyongyi and H. Garcia-Molina, "Web Spam Taxonomy". Proc. of the First International Workshop on Adversarial Information Retrieval on the Web", 2005.

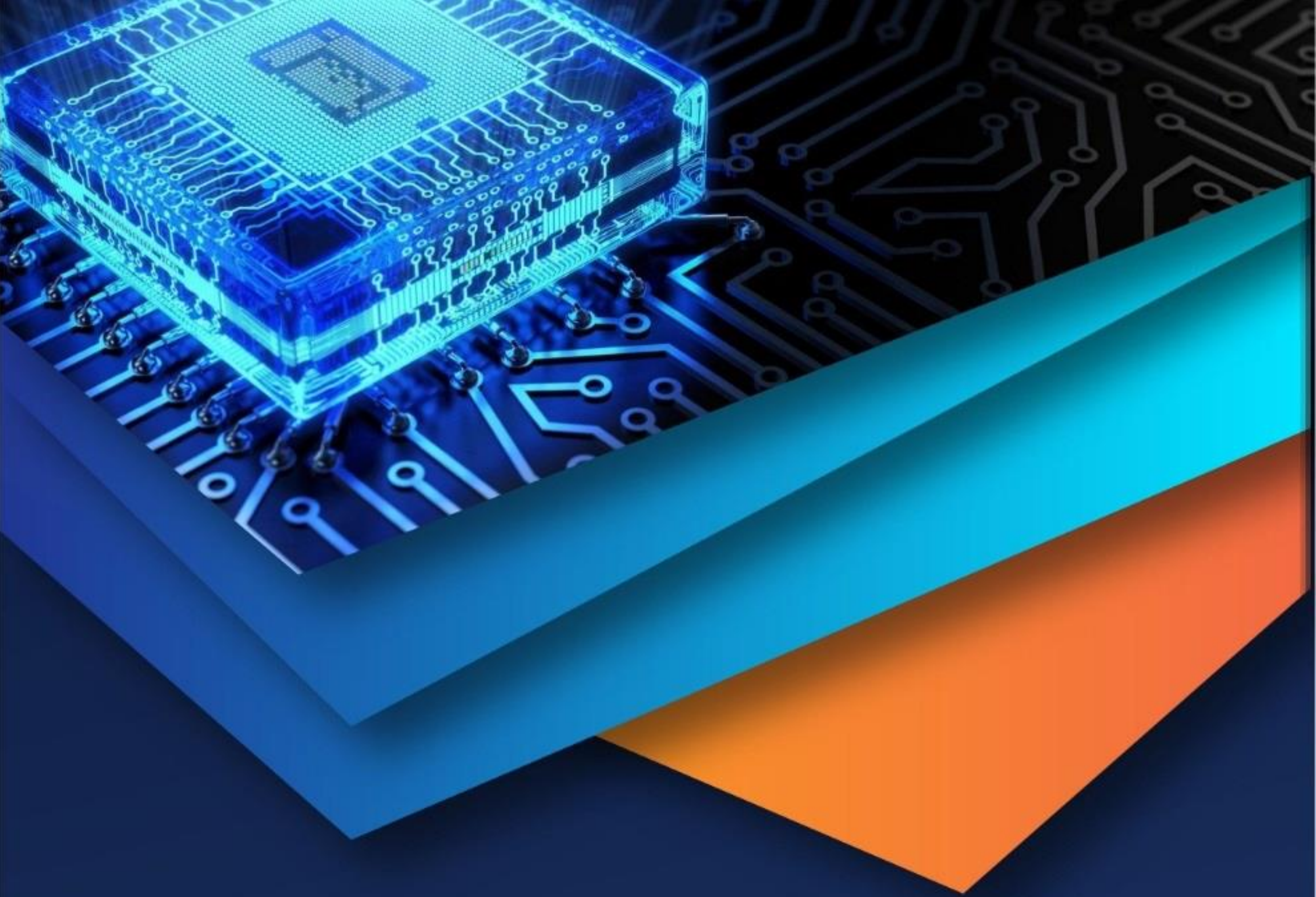
[20] M. Bianchini, M.. Gori and F. Scarselli, "Inside PageRank". ACM Transactions on Internet Technology, Vol. 5, Issue 1, 2005

[21] C.. H. Q. Ding, X. He, P. Husbands, H. Zha and H. D. Simon, "PageRank: HITS and a Unified Framework for Link Analysis". Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2002.

[22] J. Cho and S. Roy, "Impact of Search Engines on Page Popularity". Proc. of the 13th International Conference on WWW, pp. 20-29, 2004.

[23] J. Cho, S. Roy and R. E. Adams, "Page Quality: In search of an unbiased web ranking". Proc. of ACM International Conference on Management of Data". Pp. 551-562, 2005.

[24] A. M. Zareh Bidoki and N. Yazdani, "DistanceRank: An intelligent ranking algorithm for web pages" Information Processing and Management, Vol 44, No. 2, pp. 877-892, 2008.



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY