International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

Similarity Measurement Technique for Measuring the Performance of Page Rank Algorithm Based On Hadoop

M. A. H. Wadud, M. A. Jafor, M. F. Mridha, M. M. Rahman

Abstract: ***In this century big data manipulation is a challenging task in the field of web mining because content of web data is massively increasing day by day. Using search engine retrieving efficient, relevant and meaningful information from massive amount of Web Data is quite impossible. Different search engine uses different ranking algorithm to retrieve relevant information easily. A new page ranking algorithm is presented based on synonymous word count using Hadoop MapReduce framework named as Similarity Measurement Technique (SMT). Hadoop MapReduce framework is used to partition Big Data and provides a scalable, economical and easier way to process these data. It stores intermediate result for running iterative jobs in the local disk. In this algorithm, SMT takes a query from user and parse it using Hadoop and calculate rank of web pages. For experimental purpose wiki data file have been used and applied page rank algorithm (PR), improvised page rank algorithm (IPR) and proposed SMT method to calculate page rank of all web pages and compare among these methods. Proposed method provides better scoring accuracy than other approaches and reduces theme drift problem.***

Keywords: Inlink, Outlink, Keyword, PageRank, Hadoop, Iterative MapReduce, Link Analysis.

I. INTRODUCTION

Before search engine invented internet was not so easy to find anything. Because people didn't know any specific address where one's desired information have. They had to enter specific and correct web URL to find content. To solve this problem, Alan Emtage1 a student at McGill University in Montreal created first search engine named "Archie" in 1990. Archie gathered all scattered information and stored in a database then match with user query. Day by day various search engine launched. In 1998 search engine Google launched.

Manuscript published on January 30, 2020.
* Correspondence Author
M. A. H. Wadud*, Department of CSE, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh. Email: mahwadud@gmail.com
M. A. Jafor, Department of CSE, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh. Email: majafor29@gmail.com
M. F. Mridha, Department of CSE, Bangladesh University of Business and Technology, Dhaka, Bangladesh. Email: firoz@bubt.edu.bd
M. M. Rahman*, Department of CSE, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh. Email: mm73rahman@gmail.com

Google Search engine search according to user query and returned a list of web address which user did not know about these web address and easily find out specific information. The search engine Google is very popular- they use page rank algorithm in their machine learning to help process and rank information. Page rank algorithm [2s] use inlinks and outlinks to calculate page rank of a website and search engine shows the list of highest page ranked webpage. Sometimes page rank algorithm determines lowest page rank which contains user information and search engine does not shows that webpages beginning in the list. So, user do not get perfect information. There are several works on it to increase performance of page rank algorithm. Such as-

N. Tyagi and S. Sharma [3] derived an algorithm based on the inlinks and eight of inlinks through the user to rank web pages. PageRank score is determined at different values of dampening factor and only inlinks are examined.

S. Brin and L. Page [4] has introduced first page ranking algorithm based on hyperlink structure of web pages which provide the order of the web. PageRank algorithm normally used for huge amount of web pages. To rank the webpages, if one page is linked by other important pages than this page is an important page. Users don't track the straight link to reach webpage such that the user will track the appropriate link is more. Due to the fact that not all the user doesn't follow the direct link to reach the webpage so the dampening factor was also introduced into the algorithm as the probability that the user will follow the particular link. For this there is a problem of theme drift and surfer jamming in this algorithm. The dampening factor was presented into the algorithm to solve this problem.

HITS [5] algorithm is one of the major or mostly used ranking algorithm to rank a webpage. On the basis of inlinks and outlinks of any webpage HITS ranks the webpages and are incorporated to rank the retrieved webpages in HITS algorithm. The term hubs of a webpage represent the inlinks to the webpage and the term authorities represents the outlinks which are links going away from the current webpage. These Hubs and Authorities are the major facts in HITS algorithm. A Hub is considered as good if it is pointed to many other Authorities and an Authorities is considered as good if it is pointed by many other Hubs. There are two limitations of HITS algorithm and they are Theme Drift and Topic Drift. Theme Drift provides same importance to all the webpages which is later on removed in the extension to the HITS algorithm and Topic Drift assigns unequal importance to the retrieved webpages.

Page Ranking algorithm based on Visit of Link (VOL) [7] is another form of Page Ranking algorithm and it is the extension of standard Page Ranking algorithm. In this technique, the inlinks and the number of times the user visits a particular link of the webpage is considered. Depending on the user behavior the crawler return visit of inlinks is the past data as the input. The Page Ranking Algorithm based on VOL increases the relevancy of the webpages since here user input is considered and need of specialized crawler for visit count, problem of theme drift are the main limitations of this algorithm.

Weighted Page Rank is discussed by Wenpu Xing [6] and it is the extension of standard Page Ranking algorithm. Both inlinks and outlinks of the webpages votes for ranking of webpage. The weights to the links is assigned on the basis of the popularity. The number of inlinked pages and number of outlinked pages are refers to the popularity to a link. If the sum of number of inlinked pages and outlinked pages is more than the popularity is also more of that webpage. The Win(u,v) and Wout(u,v), are used to record the popularity of the webpages. Relevancy of webpages is less as compared to standard Page Ranking Algorithm, problem of theme drift [8] and surfer jamming exist which are counted as the limitations of this algorithm.

An Efficient Page Ranking Approach Based on Hybrid Model [1] by Lissa Rodrigues and Shree Jaswal is another form of page rank algorithm where they combine the concept of link-based mining and page level keyword search. Here they use a dictionary of several keyword. They parse user query and check according to dictionary. Page rank algorithm provides better result using this method than other method. But also have theme drift problem [8].

Theme drift stands for some pages not give result related to the user's query. For example, a user enters 'nice morning' but actually it will be 'good morning'. So, user can't find relevant result but SMT will not fail to provide result, it will provide result of the similar meaning of 'nice' is 'good' and user find out relevant result. In this paper, another page rank algorithm has been proposed based on similarity measurement technique to reduce theme drift problem.

The remaining of this paper is organized as follows. In Section 2, of this paper, presented a short description of the Page Rank algorithm. In Section 3, Basic introduction of Hadoop and Map-Reduce framework were discussed and proposed method has been discussed in section 4. Section 5 discusses several results based on user search query among different PageRank algorithm. Also made a comparison among three PageRank algorithms. Finally, conclusion and future directions are outlined in section 6.

II. REVIEW OF PAGERANK ALGORITHM

PageRank algorithm is exhibited by Sergey Brin and Larry Page at Sandford University which extends the citation analysis in the basis of inlinks and outlinks of a page.

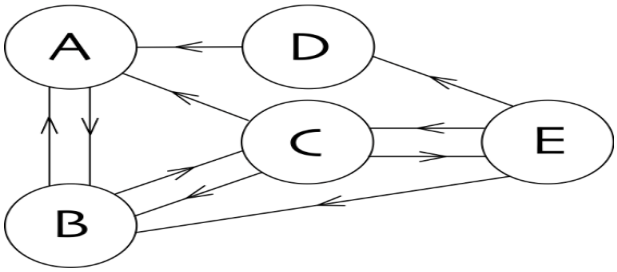

Fig. 1. Simple web page diagram connected with several inlinks and outlinks

To Web Page A, an inlink is a URL of a web page B which contains a link pointing to A. To Web Page A, an outlink is a link (URL) appearing in A which points to a web page B as shown in figure 1. The web page A which has an out-link to page B, it works as: treat B's in-links as A's votes and employ the vote to estimate the importance of A. Page and Brin proposed an equation to calculate PageRank of a page in a network containing N number of pages as below-

$PR(A) = (1-d) + d * (PR(T_i) / C(T_i))$

Where i=1,2,3…..N, $PR(T_i)$ is the rank of page T_i and $C(T_i)$ is the number of outlinks of page T_i and d is the damping factor whose value is 0.85.

III. HADOOP AND MAP-REDUCE

Hadoop MapReduce is a popular big data processing engine. Hadoop MapReduce is a software framework for distributed processing of large data sets on compute clusters of commodity hardware. Hadoop [15] was created by computer scientists Doug Cutting and Mike Cafarella in 2006 to support distribution for the Nautch search engine. After years of development within the open source community, Hadoop 1.0 became publicly available in November 2012 as part of the Apache project sponsored by the Apache Software Foundation. According to The Apache Software Foundation, the primary objective of MapReduce is to split the input data set into independent chunks [14] that are processed in a completely parallel manner.

Hadoop framework [13] is used to process extremely large data sets using a large number of nodes (computer). It is an open source Java-based programming framework in a distributed computing environment. HDFS and MapReduce are two core components of Hadoop. HDFS stands for Hadoop Distributed File System. For processing data, MapReduce programming model store data on the HDFS. Map and Reduce are two user defined functions of Hadoop MapReduce job. The Map step generate key-value pairs and writes the output to the HDFS. These values that are associated with the same key are processed by Reduce step. Reduce step feedback these values to a next call of the Map step. Hadoop require a global data structure to process iterative phases of MapReduce jobs in the graph analysis. In each iterative phase creates a new MapReduce job. The next MapReduce job is conveyed by the previous MapReduce job. The Hadoop MapReduce framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file system.

IV. PROPOSED METHOD

Link Structure based Algorithm works on link structure to rank pages. Which Signify that the page is more important and higher rank whose have more links than other pages. As a result, users can not show different outcome based on their given keyword and query. To exceed the problem of theme drift and improve accuracy a SMT approach is proposed of enhanced ratio rank [9, 16] and page level keywords. In this paper, the importance and relevance webpages are calculated by taking into account weight of in links, weight of out links and page level keywords. This shown in Fig. 02.

The proposed SMT model contains the following different blocks:

Step 1: A dictionary is created using a large number of predefined keywords

Step 2: User search queries using keyword from page repository.

Step 3: Page Repository is a large number of 1500 of wiki webpages. Link Structure explain the linking of webpages which is created after creation of the repository.

Step 4: Link Structure is created using Hadoop MapReduce framework [11] and set initial rank of each pages to 1.

Table-I: Dictionary of Keyword

S. NO.	KEYWORD NAME	SYNONYMS
01	Computer	Calculating Machine, Machines, Calculator
02	Crawler	Tractor, Bulldozer, Agrimotor
03	Entertainment	Relaxation, Diversion, Removal
04	Display	Show, Showing, Exhibition, Parade, Exposition, Spectacle, Opening, Blooming, Exposer, Reveal, Exert, Show off
05	Pollution	Contamination, Filth, Untidiness, Uncleanliness, Dirtiness, Profanation
06	Student	Pupil, Learner, Scholar, Son, Reader, Colleger, Collegian
07	Memory	Remembrance, Mind, Recollection, Retention, Record
08	Security	Security, Safety, Surety, Indemnity, Guaranty, Bail
09	Mathematics	Math's, Arithmetic, Enumeration, Algebra, Calculation, Calculus
10	Market	Bazaar, Bazar, Mart, Plaza, Fair, Bargain, Shop, Auction, Trading, Brokerage
11	Pressure	Arc, Press, Clot, Squeeze, Influence
12	Vehicle	Car, Coach, Hackery, Carrier, Van, Medium, Chariot, Calash
13	Quick	Fast, Quick, Rapid, Speedy, Fleet, Swift, Hasty, Expedition, Urgent, Early, Soon, Instantly
14	Air	Wind, Flatulence, Mania, Breeze, Dead Wind, Climate
15	Close	Near, Familiar, Thick, Approaching, Imminent, Intimate, Bosom
16	Open	Bare, Uncovered, Naked, Bleak, Untied, Unhindered, Frank
17	Disease	Illness, Unhappiness, Displeasure, Grief, Misfortune
18	Disaster	Upheaval, Woe, Turnover, Loss, Casualty, Lesion, Ruin
19	Search	Investigation, Query, Exploration, Inquiry, Quest, Pursue
20	Restaurant	Coffee house, Cafe, Tavern, Eating House, Dining Room

Step 5: Calculate inlinks and outlinks [12] using following eq^{n} 1 and 2

$$W_{(a,b)}^{in} = \frac{I_b}{\sum_{p=R_{(a)}}^{I_p}} \quad (1)$$

Where,

I_b = number of inlinks of webpage u
I_p = number of inlinks of webpage p
R(v)= Reference page list of v

$$W_{(a,b)}^{out} = \frac{O_b}{\sum_{p=R_{(a)}}^{O_p}} \quad (2)$$

Where,

O_b= number of outlinks of webpage
O_p= number of outlinks of webpage p
R(v) = Reference page list of v

Step 6: Table I shows dictionary list of different word with synonymous word.

Step 7: Calculate Total Frequency of each Keyword

$$Tf_i = \sum_{p=0}^{N}\sum_{j=0}^{M} SimWord[i][j] \quad \text{for i=0,1,2…n} \quad (3)$$

Step 8: Calculate Average Keyword Count (AKWC)

$$AverageF = \frac{\sum_{i=0}^{n} Tf_i}{n} \quad (4)$$

Step 9: calculate weight of each page

$$Weight_p = \frac{\sum_{i=0}^{n}\sum_{j=0}^{M} SimWord[i][j]}{AverageF} \quad (5)$$

Step 10: Apply the proposed algorithm as in following Equation [8]

$$RR(u) = (1-d) + d * KW\Sigma_{v\in B_{(u)}} \frac{V_u * .7 * W_{(v,u)}^{in} + .3 * W_{(v,u)}^{out} RR(v)}{TL_{(v)}} \quad (6)$$

Where,

RR(u) and RR(v) = ranking of the webpages u and v respectively
d = dampening factor
V_u= number of visits of link which points from v to u
$TL_{(v)}$ = Total number of visits of all links present on v
$B_{(u)}$= pages which points to webpage u
KW =keyword weight $Weight_p$ of page P
$W_{(v,u)}^{in}$ = weight of inlinks of connecting page v and u

$W_{(v,u)}^{out}$ = weight of outlinks of connecting page v and u

Step 11: Pages are arranged in ascending order of scores

In Figure 2, internet is the major medium of web page collection. Google, Yahoo, Bing etc. search engine collect huge numbers of web pages from internet and stored into page repository a webpage database. In proposed model initially set page rank value to 1 of all web pages. Then calculate all inlinks and outlinks for a specific web page. Next there have two major modules. One module is keyword matching module and another is analysis module.

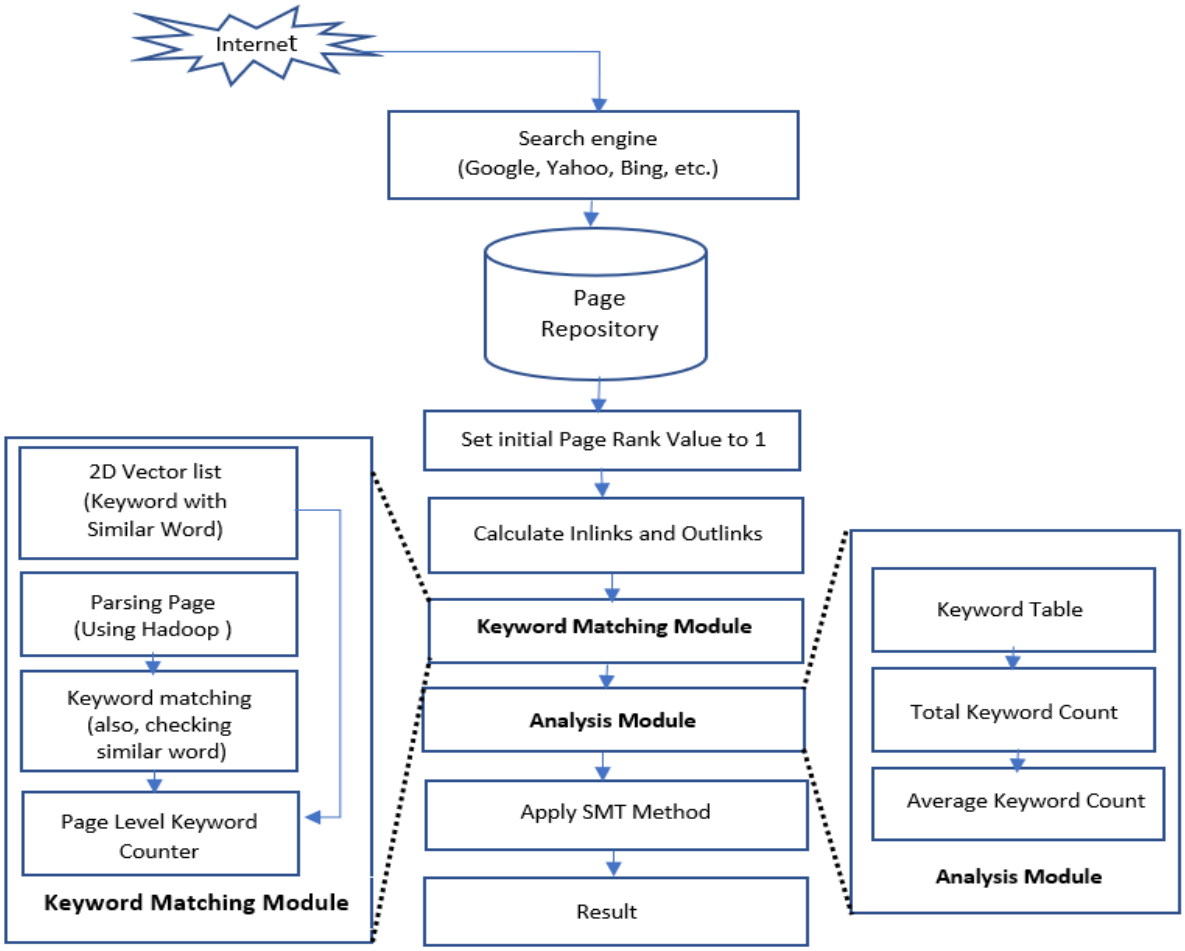

Fig. 2. Proposed SMT Model

Keyword matching module have four parts. At first there have a dictionary which contains possible list of all keyword and corresponding synonyms of each keyword respectively as like as table 1. Then Hadoop MapReduce framework parse all web pages and return keyword list with number of times a keyword appeared in the web page. Keyword matching is another method to matches all keyword and check their synonyms to find out perfect list of all keyword.

In analysis module there have also three parts. Keyword table is the list of all keyword calculated from keyword matching module. Keyword table have key, value pair list where key is keyword and value are total number of times a keyword appears in a web page. Next part count total frequency of keyword in the table and also calculate average keyword named AKWC for every web page. Then calculate percentage of keyword as keyword weight appearance in each web page based on average keyword.

Finally, proposed SMT model has been applied (as shown in eq^{n} 6) after calculating all necessary parameter. In SMT model multiplied 0.7 with inlinks weights and .3 with outlinks weights for better performance. Damping factor (d) value remain unchanged in propose model as like as original page rank algorithm which is 0.85.

Last section of proposed SMT model is result of web page rank. There considered top 10 web pages from a large collection of web page and highest ranks shows as a first web page in the list.

In SMT model main focus was reducing theme drift problem where page ranks algorithms provide same rank of different web pages which are still different categories among them. Proposed algorithm reduces the problem of theme drift which is the major problem in original page rank algorithm and shows different ranking result based on the content of corresponding web page. The new parameter Keyword KW is used to take users query and retrieved efficient and relevant result as per user's query.

V. RESULT AND DISCUSSION

For testing purpose, ten thousand websites have been taken in web database which are WIKI pages and manipulate by Hadoop [13]. Page rank algorithm (PR) [4], improvised page rank algorithm (IPR) [10] and proposed similarity measurement technique (SMT) have been applied on wiki web pages. Search results of page rank algorithm, improvised page rank algorithm was distinct from search result of similarity measurement technique and our method gives better result.

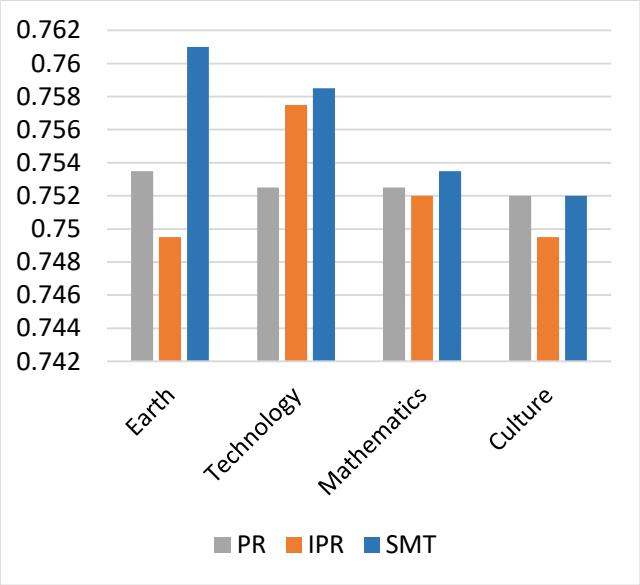

Fig. 3. Query: "Display Memory Market"

Four types of web pages such as earth, technology, mathematics and culture collected for experiment. Page repository contained these four types of thousands of web pages. Earth related all information and geographical related web pages are in earth type similarly all science, innovation and technology related web pages consider as technology type and mathematical calculation, physics etc. related web pages consider as mathematical type and people lives and their life of different region consider as culture type.

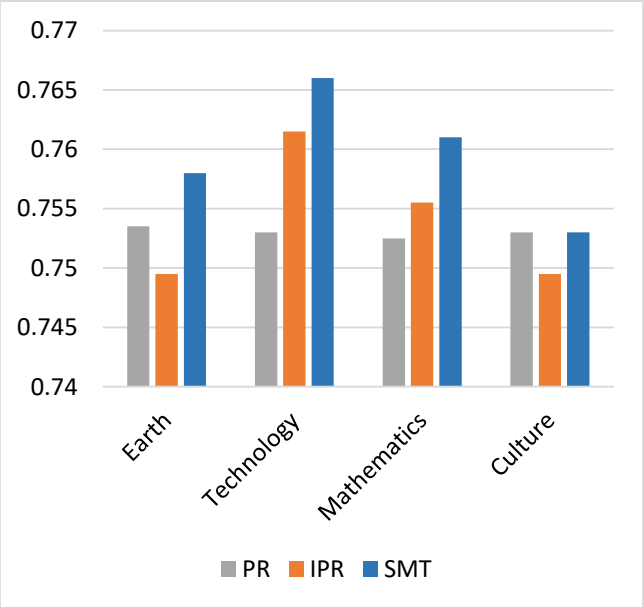

Fig. 4. Search Query: "Open Display Memory Market"

3 different cases have been applied for testing purpose. Case 1 consists display memory market search query where keywords are display, memory and market.

Case 1 have been applied into three-page rank methods such as PR, IPR and SMT model and result shown in figure 3. figure shows that SMT model produce better performance than other PR and IPR method on Earth type web pages because earth type repository consists display and market related keyword.

Similarly, in technology and mathematics type web pages proposed model shows better result than other twos. But in culture related web pages SMT model shows similar result as PR and IPR performance is poor than other two methods.

Case 2 consists open display memory market search query and result shown in figure 4, where SMT model produce better result than other two methods as previous scenario but IPR shows higher result than PR algorithm. In case 2, keywords are open, display, memory and market where open is added with case 1 scenario. There have lots of synonymous word as like as table-1 based on specific keyword. For example, Uncovered, Naked, Bleak, Untied, Unhindered, Frank are synonymous words for open keyword. Since proposed model consider all synonymous word based on certain keyword so SMT model produced better result than other two.

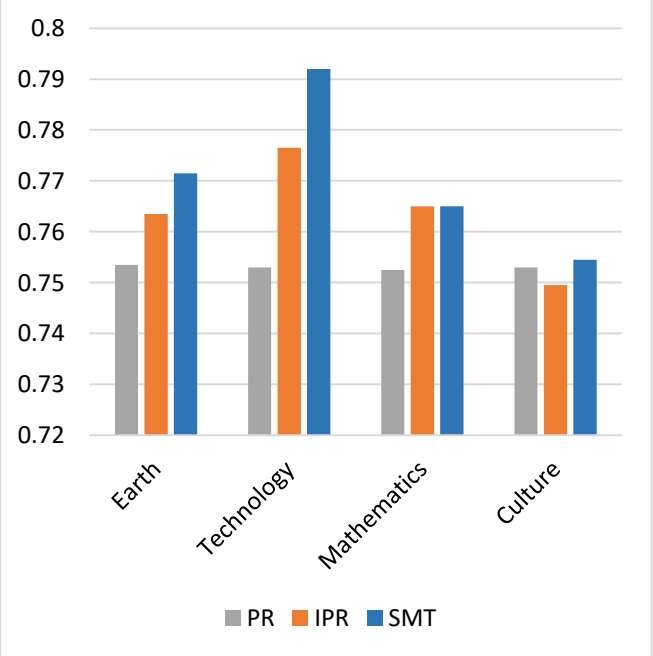

Fig. 5. Search Query: "how to setup satellite in earth orbit"

In case 3, search query is how to setup satellite in earth orbit and keywords are setup, satellite, earth and orbit. In this query SMT model produce same result as IPR for mathematical related web pages and culture related web pages produces average result but for earth and technology related web pages SMT model shows better result than PR and IPR as shown in figure 5. Comparison among the page rank algorithm, improvised page rank algorithm and similarity measurement

technique for different search query shown in figure 3, 4 and 5. Every result show that performance of proposed method is greater than existing page rank methods. In every figure SMT parse user's query in different keyword and apply synonymous keyword to calculate PageRank of web pages using Hadoop MapReduce [14] framework.

VI. CONCLUSION

In this paper, different page ranking algorithms have been experimented for finding efficient and effective results according to user's query. Original page rank algorithm, hybrid model page rank algorithm and similarity measurement technique have been implemented for comparing among web ranking algorithm. Proposed algorithm has been concluded that shows better results than others page rank algorithm. Hadoop Framework used to easily partition the web pages. This reduces the cost of merging and clustering content of web pages and increase the accuracy of experimental result. Proposed model considers keyword and similarity of each keyword for every page and then calculated rank of each pages. Pages placed on the list descending order according to their rank that means highest ranked page appeared on top of the list.

REFERENCES

1. Rodrigues, Lissa. (2015). An Efficient Page Ranking Approach Based on Hybrid Model.
2. Gupta, Renu & Shah, Ankita & Thakkar, Amit & Makvana, Kamlesh. (2016). A Survey on Various Web Page Ranking Algorithms. An international journal of advanced computer technology. 5. 7.
3. N.Tyagi and S. Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Links of Web Page", International Journal of Soft Computing and Engineering (IJSCE), July 2012.
4. Brin, & Sergey, & Page, & Lawrence,. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems. 30. 107-.
5. Jon M. Kleinberg. 1999. Authoritative sources in a hyperlinked environment. J. ACM 46, 5 (September 1999), 604–632. DOI:https://doi.org/10.1145/324133.324140.
6. Xing, Wenpu & Ghorbani, Ali. (2004). Weighted PageRank Algorithm. 305-314. 10.1109/DNSR.2004.1344743.
7. Kumar, Gyanendra & Duhan, Neelam & Sharma, Ashok. (2011). Page ranking based on number of visits of links of Web page. 11-14. 10.1109/ICCCT.2011.6075206.
8. Rutusha Joshi, Vinit Kumar Gupta, "Improving Pagerank Calculation by using Content Weight", International Journal of Science and Research (IJSR), June 2014
9. Singh, Ranveer & Sharma, Dilip. (2013). Enhanced-RATIORANK: Enhancing Impact of Inlinks and Outlinks. 10.1109/CICT.2013.6558107.
10. Jaswal, Shree. (2015). An Efficient Page Ranking Approach Based On Hybrid Model. 10.1109/ICACCE.2015.57.
11. Debbarma, Akashdeep et al. "Performance analysis of graph based iterative algorithms on MapReduce framework." International Conference for Convergence for Technology-2014 (2014): 1-6.
12. Kim, Sung Jin & Lee, Sang Ho. (2002). An Improved Computation of the PageRank Algorithm. Proceedings of the European Colloquium on Information Retrieval. 2291. 10.1007/3-540-45886-7_5.
13. Jiang, Lincheng & Ge, Bin & Xiao, Weidong & Gao, Mingze. (2013). BBS opinion leader mining based on an improved PageRank algorithm using MapReduce. Proceedings - 2013 Chinese Automation Congress, CAC 2013. 392-396. 10.1109/CAC.2013.6775766.
14. Choi, Hoon & Um, Jung-Ho & Yoon, Hwa & Lee, Minho & Choi, Yunsoo & Lee, Wongoo & Song, Sa-kwang & Jung, Hanmin. (2012). A partitioning technique for improving the performance of PageRank on Hadoop. 458-461.
15. Hadoop. http://hadoop.apache.org/
16. Kumar, Munish & Kumar, Ravinder. (2019). An Efficient Page Ranking Approach Based on Vector Norms using sNorm(p) Algorithm. Information Processing and Management. 56. 1053-1066. 10.1016/j.ipm.2019.02.004.

AUTHORS PROFILE

Mr. M. A. H. Wadud is a Lecturer in the department of Computer Sciecne and Engineering, Bangladesh University of Business and Technology, Dhaka, Bangladesh. He received his B.Sc. and M.Sc. Engineering degree in CSE from Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. He participated several ACM ICPC programming contests during his university life. He worked on several programming platform like Java Spring & Hibernate, Android apps developments, Python NumPy, Keras etc. for big data and deep learning analysis in several software company. His Area of interest is Big Data Analysis, Deep Learning, Natural Language Processing, Internet of Things and Machine Learning. He published a paper in a reputed journal on Network Security & IOT related filed.

Mr. M. A. Jafor received his B.Sc. and M.Sc. Engineering degree in Computer Science and Engineering from Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. He participated several ACM ICPC programming contests during his university life. His is now working as Lecture in a reputed College. His research interest includes Network Security, Big Data Manipulation and Data Mining.

Dr. M. F. Mridha is currently working as an associate professor in the department of Computer Science and Engineering of the Bangladesh University of Business and Technology. He received his Ph.D. in Computer Engineering, Jahangirnagar University (JU) in 2017. His research interests include Artificial Intelligence (AI), Machine learning, Natural Language Processing (NLP) and Universal Networking Language (UNL). He is currently working on Machine Learning (ML), Bangla Language Processing (BLP), Solving Ambiguity of Bangla words and supervised and unsupervised learning. For more than 10 (Ten) years, he is working with the master's and undergraduate students as a supervisor of their thesis works. He has served as a reviewer of various IEEE conferences like ICCIT, IJCCI, ICAEE, ICCAIE, ICSIPA, SCORED, ISIEA, APACE, ICOS, ISCAIE, BEIAC, ISWTA, IC3e, ISWTA, CoAST, icIVPR etc.

Dr. M. M. Rahman is a professor in the department of Computer Science and Engineering, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh. He Completed his Ph.D degree from Jahangirnagar University, Bangladesh. His research interests include digital image processing, medical image processing, computer vision and digital electronics. He has many international journal and conference publications.