ISSN:0975-9646

Ranking Technically Influential Users on Web Blogs

Amruta K. Jadhav[#1], Sharvari Tamane [#3]

[1]*Deaprtment of Computer Science and Engineering, Jawaharlal Nehru Engineering College, Aurangabad, Maharashtra, India*

[2]*Department of Information Technology, Jawaharlal Nehru Engineering College, Aurangabad (M.S.) India*

Abstract- The ever-increasing growth of the Internet is leading to its widespread use for various purposes. Research shows that, due to low publication barrier to content uploader, anonymity of uploader, exposure to millions of users and a potential of a very quick and widespread diffusion of message, various Internet platforms like blogs have become one of most important parts to promote our ideologies in a sophisticated manner. This, together with growing popularity of online blogs, calls blog service provider for providing the relevant and quality information to the web user against their query.
We experiment here approaches to rank technical users on blogs with the help of Naïve Bayes classification algorithm and Google page rank algorithm. Evaluation on real-world data from online web blogs is done to determine which algorithm is performing best. The result shows that the performance obtained by the Google page rank algorithm is better than that achieved by Naïve Bayes classification algorithms

Keywords- **Web content mining, ranking, ranking algorithms, blog, discussion forums Introduction**

I. Introduction And Motivation

Over the past decade, online discussion forums, like blogs have emerged into a dynamic form of world-wide interpersonal communication. As the volume of information on the internet is increasing day by day there is a challenge for blogging service providers to provide proper and relevant information to users.

Due to simplicity of navigation, low publication barriers (users only need to have a valid account on website) and anonymity of content uploader (liberty to upload any content without revealing their real identity) have led users to misuse blogging services in several ways by uploading spam or irrelevant data [1].

Dynamically increasing unstructured or semi-structured information on the blogging websites lead a great challenge for both the users, who are seeking for efficiently valuable information and for the technical people, who are offering service to an individual user, covered in the billions of blog posts. To triumph over these problems, data mining techniques must be useful on the www. Nowadays, most of the people rely on web search engines to find and retrieve information [29]. The enormous growth, assorted, dynamic and unstructured nature of web makes internet awfully difficult in searching and retrieving relevant information and in presenting query results. Every day search engines are giving response to millions of queries. An efficient ranking of query words has a major role in efficient searching for queried content. There are various challenges associated with the ranking of blog posts such that some blog posts are made only for navigation purpose and some of the posts do not possess the quality of self descriptiveness. Over past 10 years, since 2005 several approaches, techniques, algorithms and tools have been proposed for ranking of web pages and to bring solutions to detect blog users based on their behavioral features.

A. Role of Influential Users

Due to enormous and rapid growth of user-generated content on web blogs, a significant portion of such data remains just a noise, and users generally avoid going through every comment posted by others. There always exist some users who develop some trust relationships with other members by their activeness and quality of comments, and their comments always receive significant amount of attention among online community. These are the influential users, who play a leading and dominating role in the web blogs, and their activities and comments greatly affect the sentiments of others [9]. For example, the popularity of a technical blog is completely dependent on the owner's influence, where a majority of users remain silent spectators following the few influential technical leaders. As a result, be it a product campaign or product marketing or technical ideology propagation, influential users most of the time find it very easy to convince the silent spectators and promote their ideologies.

B. Need for Ranking

There are billions of web pages, blog posts on the web and it is more than likely that when a user enters a word to be searched for there will be thousands of results containing that word. It is obviously impractical for the user to visit all of these pages. Thus, one of the goals of a search engine is to provide the user with results that are most likely to be beneficial to him/her in least possible amount of response time.

When the search engines return the result of a user query, only a predetermined number of documents are presented to the user. Thus, it is very important that the most relevant documents are included in the result and are prioritized in the display. This important task is performed by the ranking function. A ranking function that prioritizes the documents most relevant to a user will satisfy the user.

C. Our Contribution

We make the following key contributions in this paper. i) An application of Naïve Bayes classification algorithm and Google page ranking algorithm to rank technical users on web blogs. ii) A measure to compute the degree of technicality of a user based on the degree of match his/her posts with a manually crafted list. iii) An experimental analysis of a real-world web blog dataset and to define users' technicality and to calculate technicality score for

each users. iv) An empirical evaluation of algorithms with real-world data sets that each has different characteristics.

The rest of the paper is organized as follows. Section II presents a review of the related works, followed by proposed system in Section III. Section IV presents the proposed method, and finally, Section V concludes the paper with few important future research directions.

II. RELATED WORK

As the number of blogging sites is increasing regularly, there is a challenge for service provider to provide good blogs to the users. An Eigen Rumor algorithm [7] is proposed for ranking the blogs. This algorithm provides a rank score to every blog by weighting the scores of the hub and authority of the bloggers depending on the calculation of Eigen vector. This algorithm enables a higher score to be assigned to the blog entries submitted by a good blogger but not yet linked to by any other blogs based on acceptance of the blogger's prior work.

Web page developers give more importance to some links using different HTML tags, because some Web resources are more significant than others. Hence, a link ranking technique that gives different weights to links may improve over uniform weight links [8]. This algorithm provides weight value to the link based on three parameters i.e. length of the anchor text, tag in which the link is contained and relative position in the page. Simulation results show that the results of the search engine are enhanced using weighted links.

An innovative algorithm named as Tag Rank [6] is proposed by Shen Jie, Chen, Zhang Hui, Sun Rong-Shuang, Zhu Yan and He Kun. It deals with ranking the web page based on social annotations. This algorithm calculates the heat of the tags by using time factor of the new data source tag and the annotations behavior of the web users. This algorithm provides a better authentication method for ranking the web pages. The results of this algorithm are very precise and this algorithm index new information resources in a better way.

A query dependent raking algorithm for search engine have been presented by Lian- Wang Lee, Jung- Yi Jiang, ChunDer Wu and Shie-Jue Lee [5], where a simple similarity measure algorithm is used to measure the similarities between the queries. A single model for ranking is made for every training query with consequent document. Whenever a query arises, then documents are extracted and ranked depending on the rank scores intended by the ranking model. The ranking form in this algorithm is the combination of various models of the similar training queries. Experimental results show that query dependent ranking algorithm is better than other algorithms.

Ali Mohammad Zareh Bidoki and Nasser Yazdani [18] has proposed an intelligent ranking algorithm named as distance rank, which is based on reinforcement learning algorithm. In this algorithm, the distance between pages is considered as a punishment factor. In this algorithm the ranking is done on the basis of the shortest logarithmic distance between two pages and ranked according to them.

Fabrizio Lamberti, Andrea Sanna and Claudio Demartini [21] proposed a relation based algorithm for the ranking the web page for semantic web search engine. Various search engines are presented for better information extraction by using relations of the semantic web. This algorithm proposes a relation based page rank algorithm for semantic web search engine that depends on information extracted from the queries of the users and annotated resources. Results are very encouraging on the parameter of time complexity and accuracy.

Lian- Wang Lee, Jung- Yi Jiang, ChunDer Wu and Shie-Jue Lee [22] have presented a query dependent raking algorithm for search engine. In this approach a simple similarity measure algorithm is used to measure the similarities between the queries. A single model for ranking is made for every training query with corresponding document. Whenever a query arises, then documents are extracted and ranked depending on the rank scores calculated by the ranking model. The ranking model in this algorithm is the combination of various models of the similar training queries. Experimental results show that query dependent ranking algorithm is better than other algorithms.

M Vojnovic et al. [23] have proposed a ranking and suggestive algorithm for popular items based on user feedback. User feedback is measured by using a set of suggested items. Items are selected depending on the preferences of the user. The aim of this technique is to measure the correct ranking of the items based on the actual and unbiased popularity. Proposed algorithm has various techniques for suggesting the search query. This algorithm can also be used for providing tag suggestion for social tagging system. In this algorithm various techniques for ranking and suggesting popular items are studied and results are provided based on their performance. Results of this algorithm demonstrate that randomized update and light weight rules having no special configurations provide better accuracy.

Xiang Lian and Lei Chen [25] have proposed an algorithm for ranked query processing in uncertain databases. Uncertain database management is used in various areas such as tracking of mobile objects and monitoring of sensor data. To remove these limitations authors have proposed a novel algorithm.

Tarique Anwar et. al. [2] have proposed an approach to identify a ranked list of radically influential users in Web forums, by formulating a radicalness measure and a variety of collocation-based association measures, and designed an algorithm based on Page Rank to rank the radically influential users. The experimental results on a standard data set are promising that outperforms the existing User Rank algorithm in which the contingency coefficient measure is found as the most promising measure. The result confirms that collocation-based association measures deal with such ranking problem more effectively than textual and temporal similarity based measures.

Blogs not have so efficient search engines for them. One reason is differences between regular web pages and blog pages and inefficiency of conventional web pages ranking algorithms for blogs ranking. There are some works in the field but users' behavioral features have not considered yet. M. A. Tayebi et. al. [28] presents a new

blogs ranking algorithm called B2Rank based on these features.

III. PROPOSED SYSTEM

The proposed method starts with crawling and preprocessing the web blog data, followed by technicality identification, and finally ranking the technical users based on a Naïve Bayes classification and Google page ranking algorithm respectively.

A. Forum Crawling and Preprocessing

The process starts with a data crawling and preprocessing step in which the URL of the web blog page is passed to the forum crawler, which crawls all relevant web blog pages and eliminates the duplicates. A parser module is employed to extract the meaningful snippets from the crawled web blog pages, which are then passed to the data preprocessing module. The metadata extraction task works in close coordination with the parser module to extract all relevant metadata. The obtained data is organized as a collection of threads having a unique id and title; the body text is additionally processed through some cleaning and chunking mechanisms to remove the noise and crystallize into individual meaningful pieces of information.

B. Measuring Technicality

The foundation of automatic technical user identification process is laid on a set of manually crafted list of technical words that are typically found in techno oriented texts. In some studies, the researchers manually crafted the list of technical words as a subset of the pruned list of words from the technical web blogs, which consists of English words. The forum is believed by many people as representing technical ideology. We noticed that the technical word list is quite long, and most of the words in the list are also used in general situations. Because the list is manually crafted, there needs to be strong rationality to use the words for characterizing technicality. All the words in the list except a few like support, clearly express the sense of technicality, and the exceptions, although pose a non-technical sense in usual cases, but in the context of technicality they stand for a specific meaning. In real situations, it is very likely that the potentially technical users avoid using the obvious technical terms and prefer using some other form of words. Also the terms could be acronyms or synonyms or in different languages. To handle these real scenarios, the list needs to be updated regularly with time. Shorter lists may give some technical users a chance to evade, whereas longer lists (including some general terms that are perhaps also technical in a sense) may mark even normal users as technical users. Therefore we have been extreme careful while preparing and updating the list.

C. Finding Technical Users

The proposed approach refers to implementation of Naïve Bayes algorithm and Google page ranking algorithm one after one. So, in this sub-section, algorithm details for proposed solution approach are explained.

i) Naïve Bayes Algorithm

The Naïve Bayes model involves a simple conditional independence assumption, i.e. given a class which may be positive or negative; the words are conditionally independent of each other. This assumption doesn't much affect the accuracy of text classification but makes really fast classification applicable for the problem.

Inputs to this algorithm are a seed (a user) *u*, n-gram value *Ng*, bag-of-words. Each training profile is compared with all technical word values in bag-of-words and their likelihood score for each seed calculated.

In this case, the Maximum Likelihood Probability (MLP) of words x_i belonging to a particular class c is given by (1),

$$P\left(\frac{x_i}{c}\right) = \left(\frac{Count\ of\ x_i\ in\ documents\ of\ class\ c}{Total\ number\ of\ words\ in\ documents\ of\ class\ c}\right) \ldots(1)$$

Hash tables are used to store the frequency counts of words during the training phase itself. According to the Bayes Rule, the probability of a particular user u belonging to class c_i is given by (2),

$$P\left(\frac{c_i}{u}\right) = \left(\frac{P(c_i)\,P\left(\frac{u}{c_i}\right)}{P(u)}\right) \ldots (1)$$

ii) Google Page Rank Algorithm

It is generally not practical that a subset of users exists as technically influential and others not; rather it is like a property that exists in every user with varying intensities. Therefore, here we consider the problem of identifying technically influential users as a ranking problem. Both the individual properties of technicality and influence in a user are very much regulated by the other users with whom the former interacts, in addition to one's own default properties. Therefore, the interaction linkages act crucially to determine the overall magnitude. For this nature of the influence ranking problem, some previous works found the concept of Page Rank algorithm as much suitable to establish its foundation [2], [18], [21].

The Google page rank algorithm computes a ranking of web pages to find their probable importance to Web navigators and page authors. The authors generally hyperlink the important terms to refer the detail in other WebPages. It considers these Web hyperlinks as recommendations made by the directing page for the page to which the former is linking. To compute the ranking score of webpage, each of them is initialized with a small value as their page rank score (PR(p_i)), and the linkages (L) among them are iteratively used to compute their new page rank score (PR(p_j)).

In this approach, threaded discussions among users in a web blogs are used to construct a directed graph by adding each user in the web blog as a node, and each user interaction as a directed link. Unidirectional links from all commenter's to the thread initiator and bi-directional links between each pair of commenter's are established for each thread in the graph Each user node is initialized with a small value as its page-rank score, and just like the Page Rank algorithm, the directed linkages among them are used iteratively to keep on updating their rank scores, until a convergence is achieved.

IV. Experimental Results

In this section, the effectiveness of Naïve Bayes classification and Google page ranking algorithm has been investigated. This investigation is done by applying our algorithms on real-world dataset. In this section, we compare the performance of Google page rank algorithm with that of Naïve Bayes classification algorithm.

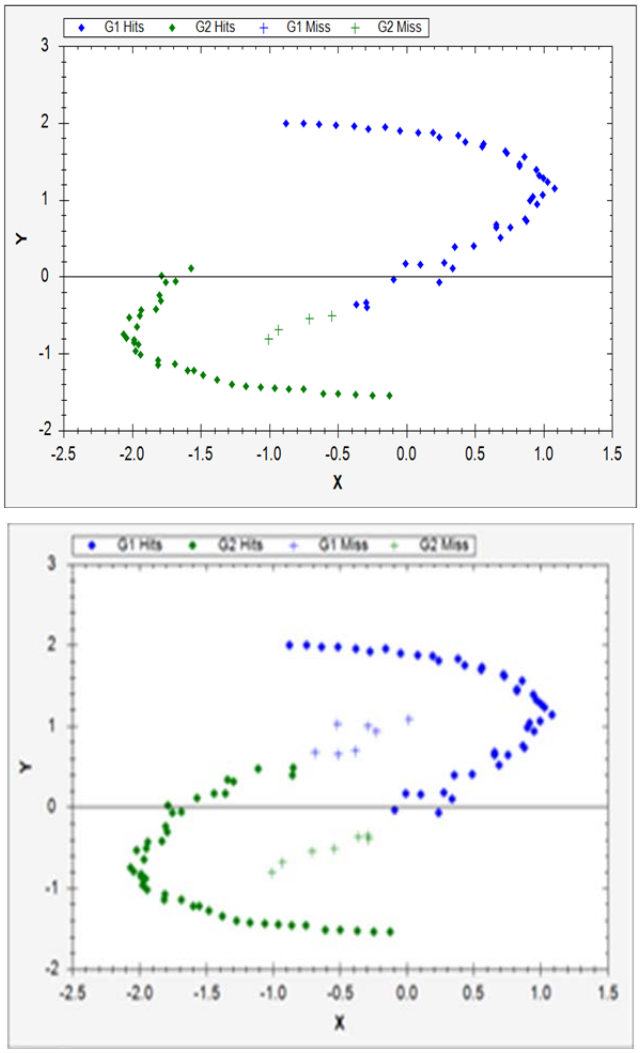

Figure 1: Visualization of Performance Measures (a) Naïve Bayes Classification (b) Algorithm Google Page Ranking Algorithm

Real-world dataset for web blogs are used in the experiments we followed. The randomly selected training documents are used for training or validation and the testing documents are used for testing. Whereas, the data for training or validation are separate from the data for testing in each case.

In this empirical analysis, we compare the performance of said two algorithms. The performance is evaluated by the accuracy, AC, which compares the predicted label of each document with that provided by the document corpus. Table I shows the accuracy results for proposed approaches and Figure 1 gives the visualization of performance measures.

Table 1: Accuracy results

Algorithm	Naïve Bayes	Google Page Rank
True Positives	36	43
False Negatives	10	7
True Negatives	46	43
False Positives	8	7
Sensitivity	0.7826	0.86
Specificity	0.8518	0.86
Efficiency	0.8172	0.86
Accuracy	0.82	0.86

V. Conclusion

In the current era, the user always desires to get the best in a petite time. User generally spends a lot of time in sifting through the search results to find the relevant information. The ranking algorithms, which are significance of web mining, play a major role in making the user search navigation easier in the results of a search. Solutions to rank web documents and blog posts on the Internet have recently attracted a lot of research attention.

In this paper, we have proposed an approach to identify technical users in web blogs. We have implemented Naïve Bayes algorithm and Google Page Rank algorithm, separately, to rank technically influential users. Among the implemented algorithms, the Google Page Rank algorithm is found as the most promising algorithm. The experimental results on a real-world data set are promising that outperforms and confirms the accuracy of Google Page Rank algorithm (86%).

VI. Future Work

This work opens several promising directions for future research. Considering social relations in addition to the threaded interactions, exploring semantic factors like discussion context and topic drift for technicality identification, and applying sentiment analysis to differentiate between the users taking positive and negative sides are few important research problems.

The algorithms and the data sets adopted are intended to be popular and easily accessible for anyone interested in this research area. However, it would be of greater value evaluating the performance of the measures on larger test-beds. Also, this work mainly focuses on textural features. It would be interesting to investigate the effectiveness and efficiency in the scenarios that involve non-textual features and objects.

Acknowledgement:

One of the authors (Amruta Jadhav) is thankful to Dr. (Mrs.) Madhuri Joshi, Head, Dept of Computer Science and Engineering, and Dr. S. D. Deshmukh, Principal, Javaharlal Nehru Engineering College, Aurangabad, Maharashtra, India for kind cooperation, fruitful discussions and suggestions to carry out this work.

REFERENCES

[1] Muthiah, Sathappan, et al. "*Planned Protest Modeling in News and Social Media*." *AAAI*. 2015.

[2] Tarique Anwar, Muhammad Abulaish, "*Ranking Radically Influential Web Forum Users*," IEEE Transactions on Information Forensics and Security, vol. 10, no. 6, pp. 1289-1298, 2015

[3] G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma, K.Sarukesi ,"*Relevance Ranking and Evaluation of Search Results through Web Content Mining*", International Multi Conference of Engineers and Computer Scientists 2012 vol.1,IMECS 2012.

[4] Pooja Sharma et al.," *Weighted Page Content Rank for Ordering Web Search Result*", International Journal of Engineering Science and Technology vol. 2, no. 12), pp. 7301-7310, 2010.

[5] Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, Shie-Jue Lee, "*A Query-Dependent Ranking Approach for Search Engines*", Second International Workshop on Computer Science and Engineering, vol.1, PP. 259-263, 2009.

[6] Shen Jie, Chen, Zhang Hui, Sun Rong-Shuang, Zhu Yan and HeKun, "TagRank: *A New Rank Algorithm for Webpage Based on Social Web*" In proceedings of the International Conference on Computer Science and Information Technology, 2008.

[7] Kos Fujimura, Takafumi Inoue and Masayuki Sugisaki,, "*The Eigen Rumor Algorithm for Ranking Blogs*", 2nd Annual Workshop on the Weblogging Ecosystem, 2005.

[8] Ricardo Baeza-Yates and Emilio Davis ,"*Web page ranking using link attributes*" , 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329,2004.

[9] Wenpu Xing and Ali Ghorbani, "*Weighted Page Rank Algorithm*", 2nd Annual Conference on Communication Networks & Services Research, pp. 305-314, 2004.

[10] D. Cohn and H. Chang, "*Learning to Probabilistically Identify Authoritative Documents*", 17th International Conference on Machine Learning, pp. 167–174, 2000.

[11] 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), PP.60–67, 1999.

[12] Kleinberg, Jon M. "*Authoritative sources in a hyperlinked environment.*" Journal of the ACM (JACM*)* 46.5 (1999): 604-632.

[13] Jon Kleinberg, "*Authoritative Sources in a Hyperlinked Environment*", Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.

[14] Wenpu Xing and Ali Ghorbani, "*Weighted PageRank Algorithm*", 2nd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

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

[16] D. Cohn and H. Chang, "*Learning to Probabilistically Identify Authoritative Documents*", 17th International Conference on Machine Learning, pp. 167–174, 2000.

[17] Sung Jin Kim and Sang Ho Lee, "*An Improved Computation of the PageRank Algorithm*", European Conference on Information Retrieval (ECIR), 2002.

[18] Ricardo Baeza-Yates and Emilio Davis,"*Web page ranking using link attributes*", 13th International World Wide Web conference on Alternate track papers & posters, PP. 328-329, 2004.

[19] Ali Mohammad Zareh Bidoki and Nasser Yazdani, "*DistanceRank: An Iintelligent Ranking Algorithm for Web Pages*", Information Processing and Management, 2007.

[20] H Jiang et al., "TIMERANK: *A Method of Improving Ranking Scores by Visited Time*", Seventh International Conference on Machine Learning and Cybernetics, 2008.

[21] Shen Jie,Chen Chen,Zhang Hui,Sun Rong-Shuang,Zhu Yan and He Kun, "TagRank: *A New Rank Algorithm for Webpage Based on Social Web*", International Conference on Computer Science and Information Technology,2008.

[22] Fabrizio Lamberti, Andrea Sanna and Claudio Demartini, "*A Relation-Based Page Rank Algorithm for. Semantic Web Search Engines*", IEEE Transaction of KDE, Vol. 21, No. 1, Jan 2009.

[23] Lian-Wang Lee, Jung-Yi Jiang, ChunDer Wu, Shie-Jue Lee, "*A Query-Dependent Ranking Approach for Search Engines*", Second International Workshop on Computer Science and Engineering, vol. 1, pp. 259-263, 2009.

[24] Milan Vojnovic et al., "*Ranking and Suggesting Popular Items*", IEEE Transaction of KDE, Vol. 21, No. 8, Aug 2009.

[25] NL Bhamidipati et al., "*Comparing Scores Intended for Ranking*", IEEE Transactions on Knowledge and Data Engineering, 2009.

[26] Lian, Xiang, and Lei Chen. "*Ranked query processing in uncertain databases*." *Knowledge and Data Engineering, IEEE Transactions on* 22.3 (2010): 420-436.

[27] "*Ranking Algorithm*", http://orion.lcg.ufrj.br/Dr.Dobbs/books /book5/chap14.htm

[28] The Google Search Engine: Commercial search engine founded by the originators of PageRank.

[29] M. A. Tayebi, S. M. Hashemi, A. Mohades, "B2Rank: *An Algorithm for Ranking Blogs Based on Behavioral Features*", IEEE/WIC/ACM International Conference on Web Intelligence, pp. 104-107, 2007.

[30] S. Jayanthi "*Taxonomies of Web Mining*"20th February 2013 by Web mining Research