Intelligent Web Mining to Ameliorate Web Page Rank using Back- Propagation Neural Network

Dheeraj Malhotra
Department of Computer Science and Informatics
University of Kota
Rajasthan, India
dheerajmalhotra@ymail.com

***Abstract*—In a short span of time, less than 15 years, the web search process is modified enormously because of magnificent growth in web based information resources. The speedy expansion of web is enjoyable because of the increase in information resources but at the same time its huge size and interference of SEOs in search process lead to increased difficulty in extracting relevant information from the web. Personalized web search may be the solution to relevancy problem but user is reluctant in giving his personal information because of privacy concerns [1]. Moreover most existing web mining algorithms do not possess attractive time and space complexities and hence lead to sufferings of novice user. This paper addresses above mentioned issues of Search Engine domain and intends to implement intelligent web mining in the form of *Web Page Ranking Tool* so as to improve the web page ranking process through incorporation of Back Propagation neural networks. [2][3][4]**

Keywords—Web Neural Mining, Back Propagation Neural Networks, Web Page Ranking, Intelligent Mining

I. INTRODUCTION

Information seeking is one of the most natural needs of human behaviour and how it is retrieved is dependent upon number of factors varying from personal choices to technical requirements. User browsing the web indicates the importance of web pages. Most of the popular search engines lack the ability to consider the user interaction with web for document ranking. Old web mining algorithms are not capable enough to utilize the relevance implied by user surfing patterns to improve the ranking of web pages. In recent years, personalized search has attracted interest in the research community as a mean to decrease search ambiguity and erroneous web page ranking in search results. Such a search is more likely to be interesting to a search engine user to have more accurate, effective and efficient information access. With the constant development of those researches, the intelligent search technology with the feature of adaptability and learning is also transiting from laboratory stage to practical application stage. At present, search engine and artificial intelligence domain especially the neural networks have become the key technology and core idea of Internet information retrieval [2]. Such a unique combination of data mining along with neural networks may be well utilized for web page ranking such as for E- Commerce websites to assist the customers by finding relevant web pages on the top and businesses to improve their profits to remarkable extent by providing enriched and relevant content via their E Commerce websites to their customers.

II. LITERATURE REVIEW

Web Mining when supported by neural network technology is termed as Intelligent Web Mining which can well learn from errors. The objective of such a unique combination is to present search results to user which are based on user's preferences. However all the old techniques of Web mining are not efficient enough to satisfy the continuously growing needs of present day user and hence the research in this area is continuously emerging. Dheeraj Malhotra, Neha Verma [4] well described Web dictionary based page rank determination algorithm. The proposed algorithm determine relevancy of web page using page content and time spent by previous user. The objective is to improve the time and space complexities of search engine algorithms while searching in their huge web databases without compromising with user experience. Shuo Wang, Kaiying Xu, Yong Zhang, Fei Li [2] used an approach for personalization and optimization of search engine which is based on back propagation neural networks. This method is based on feedback from user which may be explicit or implicit. They also explained various methods for obtaining training samples of information retrieval. LI Yaolin, ZHONG Yanhua, NIE Shuzhi [3] suggested quantum self organizing neural network can better cluster users for dynamic generation of personalized web pages for different categories of users. The proposed model has strong generalization capability. Kun Lin [11] discussed the need of reducing the gap between data and information stored in huge web databases for taking important decisions involved in business domain as the decision makers do not have appropriate tools for extracting information from databases. Meng Cui, Songyun Hu [18] explained the concept of search engine. They discussed various types of tools and strategies for search engine optimization such as space strategy, website structure etc. G.Lidan Shou, He Bai, Ke Chen, Gang Chen [1] explained the need of PWS. They highlighted user's reluctance to disclose their private information during search has become the major barrier for PWS. They presented a client side privacy protection framework called UPS for PWS. Erick Gomez Nieto, Frizzi San Roman, Paulo Pagliosa, Wallace Casaca, Elias S. Helou, Maria Cristina and Luis Gustavo Nonato [21] proposed a visualization technique ProjSnippet to display the results of

978-1-4799-4236-7/14/$31.00 ©2014 IEEE

web queries aimed to overcome limitations such as poor ranking, failure to provide document overview etc. The proposed method employs unique layouts that may be used for optimization of snippets to preserve neighborhood structures for document overview as well as for removal of document overlap. Renyuan Wang, Yan Chen, Taoying Li, Ying Ying Yu [14] explained the page rank technology of Google and also provides overview of sorting technologies such as Alexa Ranking Technology, HillTop ranking technology. They optimized search engine ranking using Grey factor which utilizes Grey forecasting model. Sonya Zhang, Neal Cabage [17] explained SEO and important factors affecting SEOs. Debajyoti Mukhopadhyay, Manoj Sharma, Gajanan Joshi, Trupati Pagare, Adarsha Palwe [19] proposed improvements in a meta semantic search engine, Semantalli.Yuki Yasuda, Naoto Mukai, Naohiro Ishii [16] described the page rank algorithm used by Google. They also explained "*Teleportation*" agent used by page rank algorithm. Hironao YOMEGANE, Saori KITHARA, Kenji HATANO [15] proposed a rule based method for extracting the information seeking behavior using collaborative reference database developed by NDL, Japan. Andreas Kanavos, Evangelos Theodoridis, Athanasios Tsakalidis [20] explained a method that will help user to easily get relevant results corresponding to a web query. They also discussed the need for query reformulation and report the advantages of semantics in understanding the user's search interaction. Poornalatha G, Prakash S Raghavendra [12] discussed the need for reducing the latency at the client or user end. They discussed the solution to a problem of predicting next web page to be visited by the user. S. Rajasekaran [7] explained the basics of Neural Networks such as concept of neuron, various architectures of NNs, Various learning strategies, training of NN, Back propagation algorithms etc. Mohd. Wazih Ahmad, Dr. M.A. Ansari [13] discussed various challenges in design of Intelligent Information retrieval System. They also explained Soft Web Mining including application of Fuzzy, NN and Genetic algorithms for web mining. G.Poonkuzhali[10] proposed Redundancy Computation Algorithm via n X m Matrix generation and signed approach. The objective is to improve the search result in terms of precision and recall through comparison between positive and negative count of various words of user search string. Wei Xu [9] proposed a genetic algorithm to forecast unemployment rate. They also described how the GA–NN-OSS model is dominantly advantageous over other neural network models.

III. Research problem

The exponential growth in the Web has been unhidden. The Web has become the number one source of information for Internet users however the credibility of Web information appeared to have declined somewhat. Web pages are written in variety of languages and they provide information in a variety of medium including texts, animation, images, photos, sound, music and video. As the usage and size of Web is increasing, one cannot expect to search for most relevant Web page across thousands of enlisted Web pages provided by search engine in response to a search query. Moreover page ranking provided by most of the popular search engines is highly unreliable and is hugely impacted by money making businesses such as SEO which tend to show the pages of their interest on top irrespective of their content and degree of relevancy to the search query of user. There is an urgent need to draw attention towards the problem of simplifying the reliable page ranking process matching the taste and needs of user by minimizing the impact of various SEOs.

IV. Objective of Research

The overall objective of this research work is to improve the Web page ranking process by developing a *Web Page Ranking* algorithm in the form of a tool. In this research work, a neural based mathematical approach to deal with various mining problems related to time complexity is discussed and this intelligent Web mining will optimizes the use of Web dictionary, previously spend time statistic and back propagation based neural network to improve the ranking process of Web pages. The proposed algorithm will take input from number of search engines corresponding to a User search query and arrange them in correct order as per their relevancy to user. This algorithm may be merged as a module or layer in search engine to improve the web page ranking process.

V. Research Methodology

This research work utilizes neural based web content mining technique to simplify the web page ranking problem. This approach first implements a data structure in the form of dictionary to implement Web Content mining to check for relevancy of web page and hence to determine its priority. After that web page would be passed to time module which will determine priority of current page using previous user time spent statistic database. Overall Priority of web page is finally determined by using back-propagation recurrent neural network. This network has three layers Input layer, Hidden layer and output layer. Input layer accepts three inputs i.e content relevance from content module, Time spent relevance from time module and third input is User feedback. Initially it will assign random weights to these input links. The actual output of the network is compared to expected output for that particular input. This may result into an erroneous value. The connection weights in the network are gradually adjusted, working backwards from the output layer, through the hidden layer, and to the input layer, until the correct output is produced. Fine tuning the weights in this way has the effect of teaching the network, how to produce the correct output for a particular input**.** In order to determine error value, user feedback plays a key role which may be obtained in two possible ways (i) Explicit Feedback in which user is explicitly asked to report relevance judgment. (ii) Implicit Feedback in which user behavior is noted such as which document they selected for viewing, order of documents access, time spent etc. [2]. In this research work, this feedback is stored and is utilized for future judgment of same web page corresponding to similar or matching search query.

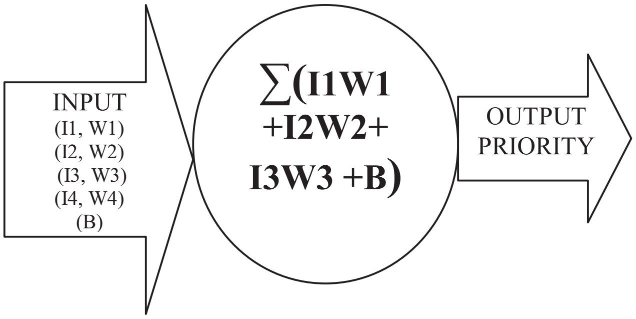

Fig.1. Simplified Neural Design of the Proposed System

As shown in above diagram, there are four possible inputs to proposed neural system i.e. Content Priority which is the priority of web page based on relevance of its content with respect to user query and is calculated from content module. This input is represented by I1. Time spent priority is second input which is the priority determined from the time spent by previous user on this page corresponding to similar or matching query. This input is represented by I2. The third input is previous user explicit feedback about the candidate web page (I3). Weights to these three inputs are represented by W1, W2, and W3. The fourth input is BIAS (B) which may be introduced as per tool requirement. These weights are assigned random values initially. These inputs and weights together will determine the output priority of web page. As the network implemented is Back-Propagation in nature so output priority is compared with desired output priority and the error is determined to adjust the weights accordingly.

A. Back Propagation Learning

Back propagation is a popular method for training of ANN (Artificial Neural Network). In this method training of an ANN involves two stages (i) Forward Stage (ii) Backward Stage. This algorithm is named back propagation because of the presence of backward phase to communicate error back to input layer. It uses Supervised Learning which employs external references to calculate the error in computed output. Some of the important steps involved in back propagation learning of proposed design are as follows [7].

Step 1: Normalize all the four possible inputs i.e. I1, I2, I3 and B represented by {I} and output priority is represented by {O}

Step2: Number of Neurons in Hidden layer are 4 represented by {H}

Step 3: Let us assume that [V] represent weight of synapses connecting Input and Hidden neurons and [W] for synapses connecting Hidden neurons and output neurons. Initialize the weights to small random values.

$[V]_{\text{initial}}$ = [random weights between -1 and +1]

$[W]_{\text{initial}}$ = [random weights between -1 and +1]

Step 4: Train the network using various set of inputs and outputs with linear activation function of

$$\{O\} = \tan Ø \{I\}$$

Step 5: Compute the inputs to Hidden layer using weight of connecting synapses as

$$\{I\}_H = [V].\{O\}_I$$

Step 6: Use Sigmoidal function for output evaluation in Hidden Layers as

$$\{O\}_H = [(1 / (1+e^{-I})]$$

Here Sigmoidal function is preferred as a non linear function as it bears more resemblance to biological neurons when compared to other activation functions and its output varies continuously.

Step 7: Compute input to output layer via multiplication of connecting synapses as $\{I\}_O = [W]\ \{O\}_H$

Step 8: Output layer will calculate output as $\{O\}_O$ using Sigmoidal function.

Step 9: Calculate the error and error rate to adjust the weights until the error rate is less than tolerance in the backward stage.

B. System Design

The proposed system consists of four modules i.e**.** Module 1: Module 2, Module 3 and Module 4. [4]

Module 1 splits user search string into words to find minimum (MIN) and maximum (MAX) length of any of the constituent words. A web dictionary from the web page is implemented by allowing only those words of web page having length between min and max.

Module 2 determines number of words of search phrase found in dictionary obtained from module 1. This module will eliminate stem words like a, an, the etc. while counting the frequency of relevant words found in candidate web page (FOUND) and also determine frequency of relevant words not found in candidate web page (NFOUND).This module will eliminate all those web pages from search results wherever FOUND is less than NFOUND.

Module 3 determine priority of each web page using previous user time spent statistic and also calculate new statistic using average calculation of current session time along with previously stored value.

Module 4 uses module 2, module 3, user feedback and back propagation algorithm to determine overall priority of web page.

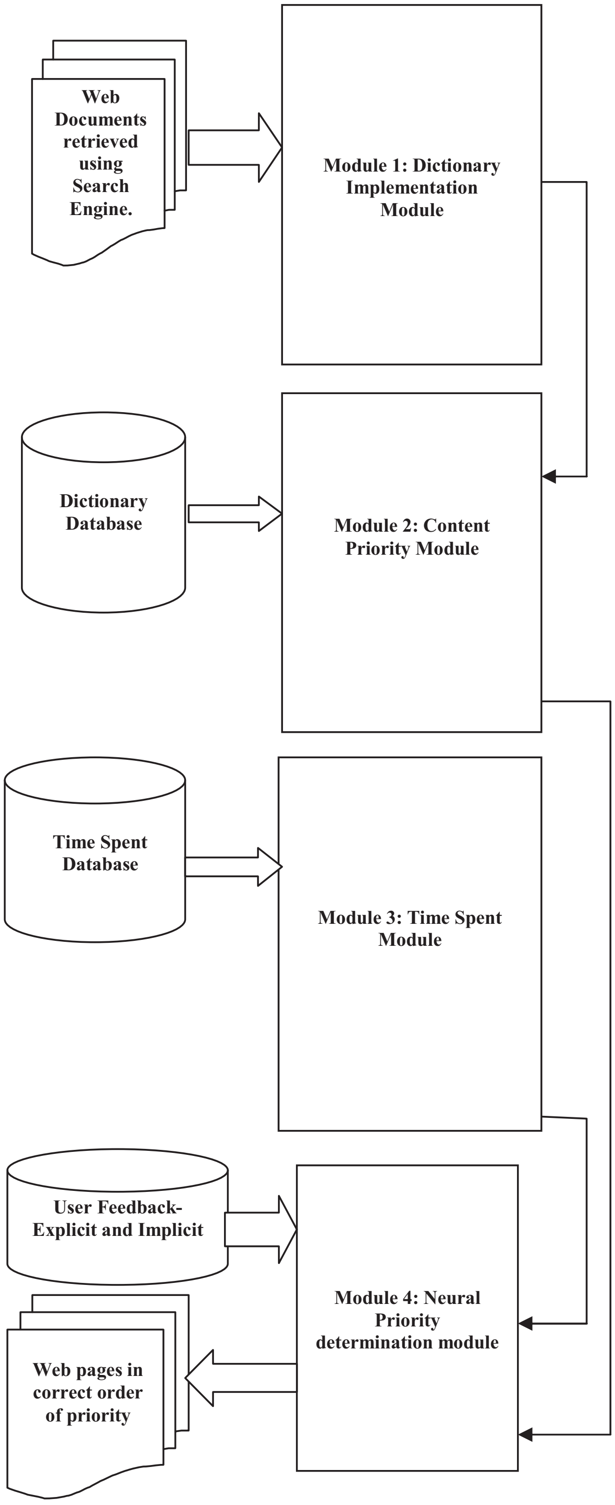
Web Documents retrieved using Search Engine.
Module 1: Dictionary Implementation Module
Dictionary Database
Module 2: Content Priority Module
Time Spent Database
Module 3: Time Spent Module
User Feedback- Explicit and Implicit
Module 4: Neural Priority determination module
Web pages in correct order of priority

Fig.2. Design of the Proposed System

C. Web Page Ranking Tool

The practical implementation of above design leads to development of a *Web Page Ranking Tool* which may be used as a layer in search engine for correct ranking of web pages. The following observation table shows the need of such a tool.

TABLE I. Sample output of Web Page Ranking Tool

Page Rank	Link	FOUND	NFOUND	Time Statistic (Seconds)	Ordered Priority
1	Link4	500	0	600	Correct
2	Link1	140	3	150	Correct
3	Link2	5	4	5	Correct
4	Link3	8	5	N/A	Correct

As shown in TABLE I, Link 4 is listed on the top because of previous user spent average time as well as count of Found variable is maximum. This tool will prioritize websites not only on the basis of content but also with reference to previous user time spent statistic, moreover where time statistic is not available (N/A), Tool is listing such link below other links where time statistic is available. Thus the proposed design can well utilize the principles of Web Content Mining [4]. The proposed algorithm may be used as a layer within search engine or may be implemented in the form of Meta Search engine. However the same algorithm may be used as a standalone tool whose interface may accept URL of number of websites along with search query and the tool will prioritize the web sites based on their relevance to search query.

D. Evaluation Metrics and Graphs

The relevancy of a webpage for a given query depends upon its position in search results. In order to compare the Tool and Google result, Precision at Y metric is considered, which is being denoted by P(Y). For a given query, P(Y) reports how many fraction of results that are labelled as relevant are reported in the top Y results. Here it is assumed that a document ranked higher is considered as more relevant and then the rank is compared with Human judgement to verify the relevancy reported by Tool and Google and finally the difference in comparison of Precision of Tool and Google is plotted. The graph shown in Fig. 3 compares the precision which is shown on vertical axis with respect to number of trial runs on horizontal axis for Google and *Web Page Ranking Tool* for the same search query. Here it is observed that the output of tool improves with repeated usage of Tool because of the fact that the back propagation neural network will adjust its weights through backward propagation of errors from output to input layer via hidden layer. This simply shows that back propagation neural networks possess the learning capabilities from errors and helps in implementation of Intelligent Web Mining.

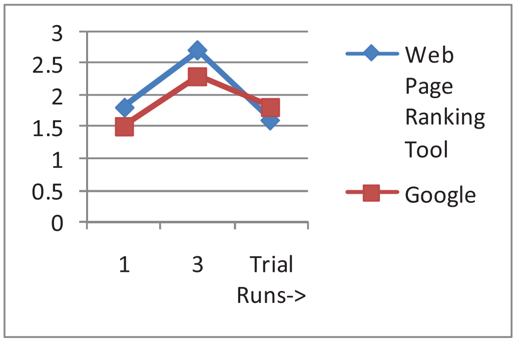

Fig.3. Precision Comparison of Tool and Google for query "Web Mining"

VI. Conclusion and Future Work

This paper presents a neural based approach for web mining to improve the search engine page ranking process. Back propagation based neural network can not only be trained better on known patterns but can also well adapt to respond to new patterns with inherent implementation of supervised learning. The capabilities of proposed system may be further improved by not just handling open web but also deep web for page ranking. Such deep web data is available through query interfaces via various authentication measures. The proposed system along with data mining technology may also be used for development of E-Commerce website optimization system. Such a website optimization system is important as latest statistics on sites like Forester clearly shows that the E-Commerce revenues of countries like India are still far less than other countries in Europe, US and even in Asia due to non efficient, non customer friendly E Commerce website structure and hence research in E-Commerce website optimization domain is urgently required and may be well supported by technologies like web mining and neural networks.

References

[1] G. Lidan Shou, He Bai, Ke Chan, Gang Chen." Supporting Privacy Protection in Personalized Web Search",IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No 2, February 2014, pp. 453-467 .

[2] Shuo Wang, Kaiying Xu, Yong Zhang, Fei Li,"Search Engine Optimization Based on Algorithm of BP Neural Networks", IEEE International Conference on Computational Intelligence and Security, IEEE Computer Society, 2011, pp. 390-394.

[3] LI Yaolin, ZHONG Yanhua, NIE Shuzhi, "Web User Access Mode Mining Based on Quantum Self organizing Neural Network", IEEE International Conference on Intelligent Computation Technology and Automation, IEEE Computer Society, 2012, pp. 382-385.

[4] Dheeraj Malhotra, Neha Verma, "An Ingenious Pattern Matching Approach to Ameliorate Web Page Rank", International Journal of Computer Applications (0975-8887) ,NewYork USA , Vol 65 , No 24, 2013, pp. 33-39.

[5] G.K Gupta,"Introduction to Data mining with Case Studies", 2nd ed., PHI 2011.

[6] J.W.Han, M.Kamber, "*Data Mining: Concepts and Techniques* ", 2nd ed. ,New York Kaufmann publishers 2006

[7] S. Rajasekaran "Neural Networks, Fuzzy Logic and Genetic Algorithms",13th ed., PHI 2010

[8] Wei Xu, Tingting Zheng, Ziang Li , "*A Neural Network Based Forecasting Method for the Unemployment Rate Prediction using the Search Engine Query Data"* , IEEE International Conference on E- Business Engineering, IEEE Computer Society, 2011, pp. 9-15.

[9] Wei Xu, Ziang Li, Qing Chen, *"Forecasting the Unemployment Rate by Neural Networks using Search Engine Query Data"* Hawaii Intearnational Conference on System Sciences, IEEE Computer Society, 2012, pp. 3591-3599.

[10] G.Poonkuzhali, G.V.Uma, K.Sarukesi, "*Detection and Removal of Redundant Web Content Through Rectangular and Signed Approach*", International Journal of Engineering Science and Technology,Vol. 2(9), 2010, pp. 4026-4032.

[11] Kun Lin,*" Applications of Web Data Mining Based on the Neural Network Algorithms in E Commerce*, IEEE International Conference on E Business Learning, IEEE Computer Society, 2010, pp. 509-512.

[12] Poornalatha G, Prakash S Raghavendra, "*Web Page Prediction by Clustering and Integrated Distance Measure",* IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012, pp. 1349-1354.

[13] Mohd. Wazih Ahmad, Dr. M.A. Ansari,"*A Survey: Soft Computing in Intelligent Information Retrieval Systems",* International Conference on Computational Science and its Applications" IEEE-CPS, 2012, pp. 26-34.

[14] Renyuan Wang,Yan Chen, Taoying Li,Ying Ying Yu, "*The Optimization of Search Engines Ranking Technology Based On Grey System",* International Conference on Computational and Informational Sciences, IEEE- CPS, 2013, pp. 1698-1700.

[15] Hironao YOMEGANE, Saori KITAHARA, Kenji HATANO, "*Extracting Information Seeking Behavior from The Collaborative Reference Database"* , International Conference on Advanced Applied Informatics, IEEE-CPS, 2013, pp. 301-305.

[16] Yuki Yasuda, Naoto Mukai, Naohiro Ishii " *Visualization of Page Rank Algorithm by Using Multi Agent Model for Education"* , International Conference on Advanced Applied Informatics, IEEE-CPS,2013, pp. 409-410.

[17] Sonya Zhang, Neal Cabage,"*Does SEO Matter? Increasing Classroom Blog Visibility Through Search Engine Optimization",* IEEE Hawaii International Conference on System Sciences, IEEE Computer Society, 2012, pp. 1610-1619.

[18] Meng Cui, Songyun Hu, "*Search Engine Optimization Research for Website Promotion"* IEEE International Conference on IT, CE and Management Sciences, IEEE Computer Society, 2011, pp. 100-103.

[19] Debajyoti, Manoj Sharma, Ganjman Joshi, *"Experience of Developing a Meta Semantic Search Engine",* International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, IEEE- CPS, 2013, pp. 167-171.

[20] Andreas Kanavos, Evangelos Theodoridis, Athanasios Tsakalidis, "*Extracting Knowledge from Web Search Engine Results"* IEEE International Conference on Tools with Artificial Intelligence", IEEE Computer Society, 2012, pp. 860-867.

[21] Erick Gomez –Nieto, Frizzi San Roman," *Similarity Preserving Snippet Based Visualization of Web Search Results",* IEEE Transactions on Visualization and Computer Graphics, 2013, pp. 1-14.