# IMPLEMENTATION OF PAGE RANKING USING GENETIC ALGORITHM

**[1]S.RAMANA MURTHY, [2]G.ANURADHA**

[1]PG –M.Tech, Dept of CSE, [2]Associate Professor, Dept of CSE,
[1, 2] GMR Institute of Technology, Rajam, Srikakulam, AP, India
E-mail: mailtomurthys@gmail.com, anuradha.govada@gmail.com

**Abstract-** Web Search engine plays a vital role in getting the information proficiently for the user needs from the immense web data. The proposed Genetic Page Rank algorithm (GPRA) based on Google's Page Rank algorithm (PRA) for the Searching purpose. Using the proposed method we are calculating the ranking for the WebPages. Genetic Algorithm (GA) is applied for ranking Web Pages in which the two parameters: Mutation is used as a similar words (synonyms) and Crossover is consider as a concept. The proposed method gives the result set for the users given query by displaying the synonym words from the Wordnet database and it ranks the webpage by taking into consideration that number of times a given keyword and their synonym words appears in that webpage as well as the hyperlink count for that keyword and also for the synonym words.

**Keywords-** Genetic Algorithm, Ontology Concepts, Ranking, Semantic Similarity, Web Search

## I. INTRODUCTION

Web mining is the application of data mining technique to discover patterns from the web. Web mining is categorized into three types: Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). Web Content Mining is used in mining of text, image, audio, video, hyperlinks and graphs of a web page to determine the relevance of the content to the search query. With the immense amount of information that is available on the World Wide Web, content mining provides the result set in order of highest relevance to the keyword in the query. Web Content Mining has different points of view: Information retrieval view and Database view. Web Structure Mining (WSM) is used to find the relation between web pages in the web structure. Web Structure Mining (WSM) can be divided into two kinds: One is extracting patterns from the hyperlinks in the web and the other is mining the document structure. Web Usage Mining (WUM) is the process of extracting useful information from server logs. It is further divided into the different kinds of used data i.e., Web Server Data, Application Server data and Application Level Data.

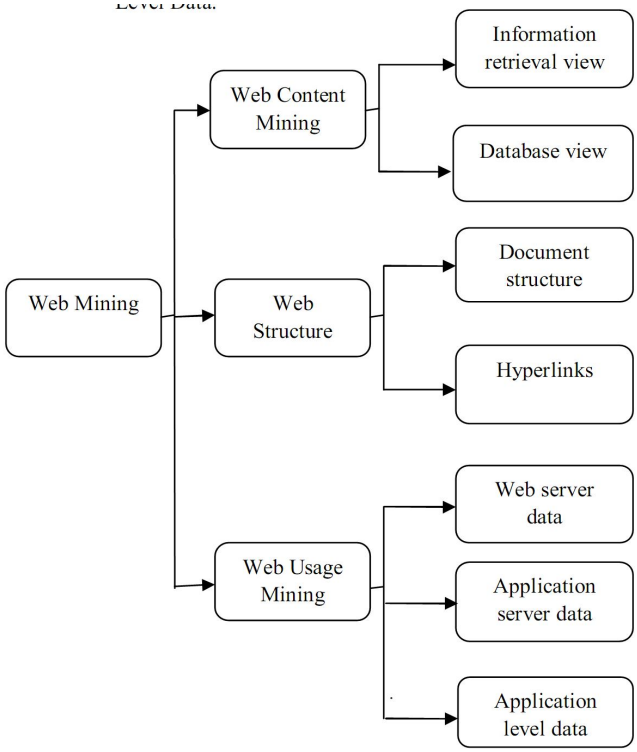



**Figure 1: Web Mining Classification**

Search engine plays an important role in getting the information competently from the huge web data. Search engine is that collects and organizes the web data. It collects the information for the user given query.

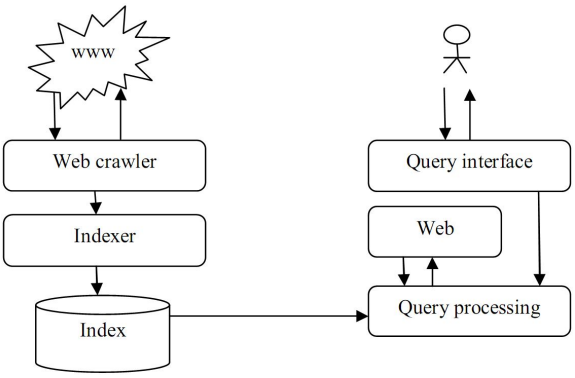



**Figure 2: Architecture of a Search Engine**

Page Rank is an algorithm used for ranking the WebPages in the search engine results. It counts the number and quality of links to a webpage to tell how important the website is. In Page Rank algorithm if a page contains important links towards it then the links of this page towards the others pages are also considered as important links. Page Rank is calculated by using equation 1.

$$PR(A) = (1-d) + d\left(\frac{PR(T1)}{C(T1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right) \quad (1)$$

- PR(A) is the Page Rank of page A,
- PR(Tn) is the Page Rank of pages Tn which link to page A,
- C(Ti) is the number of out links on page Ti
- d is a damping factor which is be set between 0 and 1.

Ranking is optimal if it retrieves all the relevant documents before any irrelevant document. From last decade Swarm Intelligence (SI) gives popular solutions for optimal problems.

## II. BACKGROUND OR RELATED WORK

PageRank: Brin and Page, et al used in the prototype of Google's search engine. The main objective is to approximate the recognition, or the significance of a webpage, based on the interconnection of the web. The motivation behind it is (i) a page containing more arriving links is more important than a page with less inward links, (ii) a page with a link from a page which is known to be of high importance is also important.

HITS (Hyperlink induced topic search): Klienberg, et al algorithm ranks the websites by processing in links and out links of web pages. In this some web pages are named as authority if that web page is pointed by many other hyperlinks and also a web page is named as hub if that web page is pointing to other web pages. In HITS algorithm ranking is determined on the structure of the web i.e., against the textual concepts for the given query.

Xing and Ghorbani, et al proposed the Weighted PageRank algorithm (WPR) that describes the importance of both the incoming links and the outgoing links of the pages and assigns rank scores based on the importance of the pages.

Baeza-Yates and Davis, et al provides weight value based on three considerations i.e., anchor text length, tag in which link is contained and relative position in the page. Length of the anchor text is one of the best attributes of this algorithm.

Fujiomura and Tanimoto, et al The Eigen Rumor algorithm is the combination of both Page rank and HITS algorithm. In this algorithm ranking is assigned for the blogs. Every blog is assigned a rank by weighting the scores of the hub and authority of the bloggers depending on the Eigen vector calculation.

Distance Rank: Bidoki and Yazdani, et al proposed a Distance Rank algorithm that looks for the “average clicks” between web pages. The main objective is to reduce the distance between pages so that a page with less distance assigned as highest rank.

Time Rank: Jiang et al., algorithm takes the first visiting time of the web page and calculates the user importance for that web pages by using the web page rank algorithm.

Tag Rank: Jie et al., This algorithm uses the time factor of the new data source tag and propose a new algorithm named TagRank based on annotations for page ranking. This algorithm gives the better authentication for ranking web pages.

## III. PROPOSED ALGORITHM

Inputs: Keyword
Outputs: Rank based web pages
Algorithm: The various steps of the proposed algorithm are given below:
Step 1: Entering a keyword in the text box of search engine.
Step 2: Calling function exec (string $command, [, array &$output [, int &$return_var ]] )
Step 3: Checks the entered keyword in the wordnet database and displays all the similar words (synonyms) for that keyword.
Step 4: con.open (); //connecting to the database using this function
Step 5: Initialize wc =0, count=0, account=0;;
Step 6: Sql command UPDATE table name SET wc= ". $count." WHERE url='". $url."'";
Step 7: Opens the url and checks whether the keyword and similar words appear in that webpage

and also checks the hyperlink count for the keyword and similar words.
Step 8: $wc=$count+$acount;
Step 9: Updates the wc value in the database depending on number of times a keyword appears in that webpage and also the similar words count as well as the hyperlink count for the keyword and similar words.
Step 10: con.close (); //Close connection from database using this function
Step 11: In the results highest ranked webpage will be displayed first.

## IV. IMPLEMENTATION DETAILS

### 4.1 Creating Database

Database consists of 426 URL dataset organized as a set of properly described table from which data can be accessed easily. The table is created by using MYSQL database which consists of following fields: sno, url, meta, desc and wc. This table consists of 64 movies related URLs, 82 health related URLs and 280 news related URLs. The meta field consists of the meta keywords with respect to the respective web pages. Desc field consists of description related to the respective web pages. This is a complete dynamic database on which several operations are performed to retrieve the highest ranked web pages that relate to the matched keyword. The following figure shows the structure of the database.

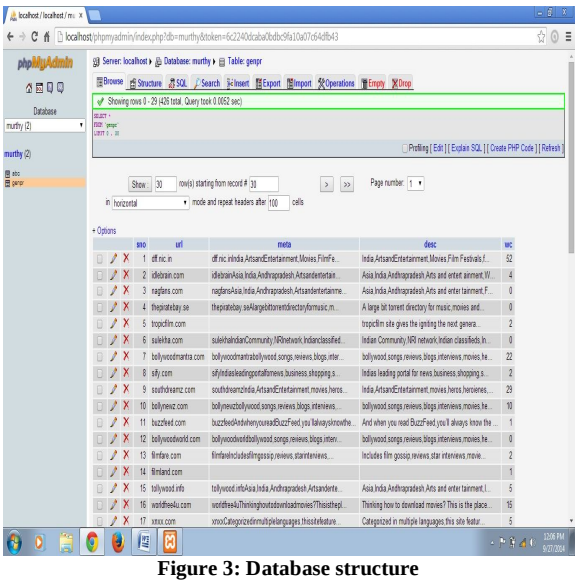


**Figure 3: Database structure**

### 4.2 User interface

After database is created in the server, user interface application of search engine is created by using PHP. It acts as an interface between the user and the search engine. In this search engine a text box is created to enter the user query as keyword and a button for searching their related links. Once the user enters the keyword in the text box of search engine, if we click on the button then it is searched in the database and the related webpage URLs are displayed.

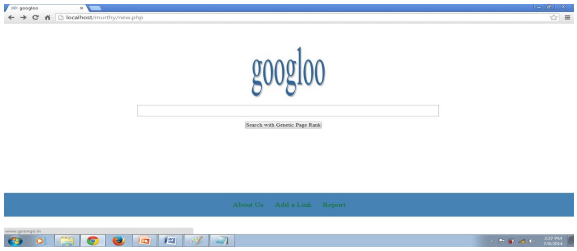



**Figure 4: Shows User interface**

### 4.3 Methodology

In the proposed method we used Genetic algorithm (GA) for ranking the web pages. Mainly we are searching results only for the three keywords namely: Movies, News and Health.GA mainly consists of two parameters: one is Mutation and the other is Crossover. Here in the proposed method we have two methods:

#### 4.3.1 Mutation:

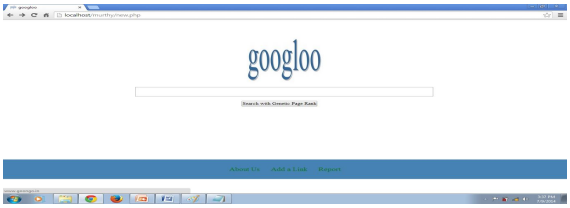
Mutation is used as similar words (synonyms). In this we used wordnet software to acquire all the synonyms for the user given query. Once the user enters the keyword in the user interface it opens the wordnet software and obtains all the synonyms for the respective keyword which are taken into an array and displayed.

#### 4.3.2 Crossover:

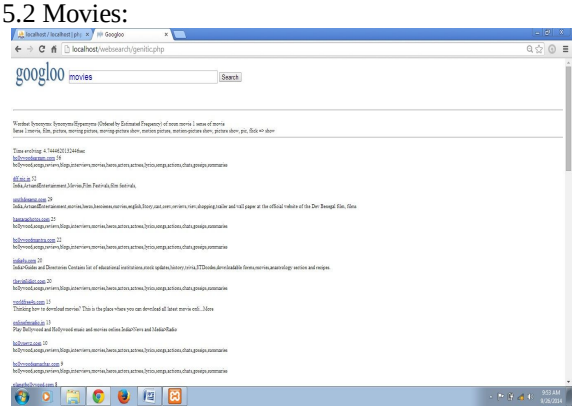
Crossover considered as a concept based ranking i.e., by taking both the frequency count and also hyperlink count of number of times a given keyword and their synonym words appears in the web page. In this proposed algorithm the WebPages with highest rank are displayed first.

## V. EXPERIMENTAL RESULTS

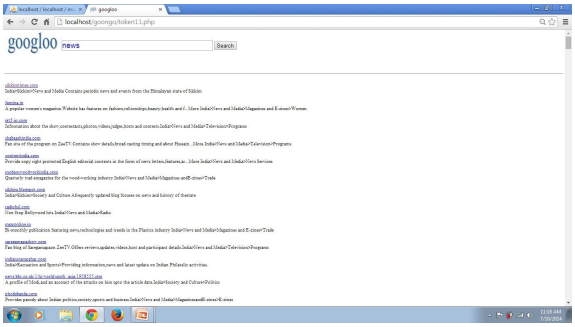
### 5.1 User interface:



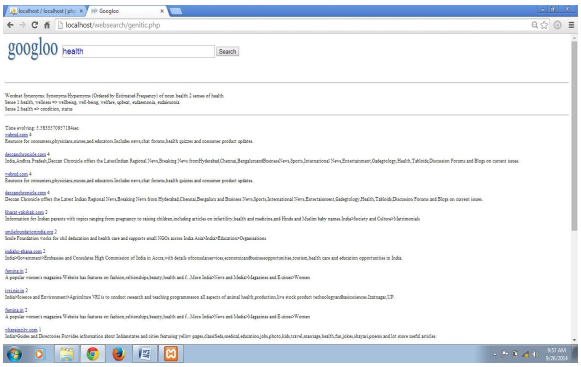

### 5.2 Movies:



5.3 News:




5.4 Health:




## CONCLUSIONS AND FUTURE WORK

On the basis of this study the conclusion is that the Genetic Page Rank Algorithm (GPRA) is based on similar words (synonyms) and concept (frequency count +corresponding hyperlink count) based ranking. The end result for GPRA returns the highest ranked WebPages for the entered query in the user interface of search engine. These results may satisfy user requirements. As a part of future guide, developed algorithms should be improved in getting the better search results.

## REFERENCES

[1] Dilip Kumar Sharma et al., “A Comparative Analysis of Web Page Ranking Algorithms” in proceedings of the International journal on Computer Science and Engineering(IJCSE), 2010.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”, Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

[3] Kleinberg J. Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999, 46(5): 604-632.

[4] Wenpu Xing and Ali Ghorbani, “Weighted PageRank Algorithm”, In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

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

[6] Ko Fujimura, Takafumi Inoue and Masayuki Sugisaki,, “The EigenRumor Algorithm for Ranking Blogs”, In WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem, 2005.

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

[8] H Jiang et al., "TIMERANK: A Method of Improving Ranking Scores by Visited Time", In proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.

[9] 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" In proceedings of the International Conference on Computer Science and Information Technology,2008.

[10] Svenson, Martenson and Micheal Malm et al., “Swarm Intelligence for logistics: Background”, 2004.

★★★