An Ingenious Pattern Matching Approach to Ameliorate Web Page Rank

Dheeraj Malhotra
Asst.Prof., Dept of IT
Vivekananda Institute of Professional Studies,
affiliated with GGSIPU, New Delhi, India

Neha Verma
Asst.Prof., Dept of IT
Vivekananda Institute of Professional Studies,
affiliated with GGSIPU,New Delhi,India

ABSTRACT

There is a spectacular growth in Web based information sources and services. It is estimated that, there is approximately doubling of Web pages every year. The rapid expansion of Web is enjoyable because of the growth of information but it is also leading to problems of increased difficulty in extracting relevant information from the Web. Most existing Web mining algorithms are not efficient enough to possess attractive Time and Space complexities. In this paper, a mathematical approach to deal with various problems related to time complexity is developed and this intelligent Web mining optimizes the use of Web dictionary and previously spend time statistic to improve the ranking process of Web pages. The proposed system can be merged as a module in search engine to improve the Web page ranking process.

Keywords

Intelligent Web mining, Pattern matching, Web page ranking.

1. INTRODUCTION

As the usage and size of Web is increasing, we cannot expect from the user to search for most relevant Web page across thousands of enlisted Web pages provided by search engine in response to a search query. In order to simplify the reliable page ranking process matching the taste and needs of user and to minimize the impact of various SEOs, there is an urgent need to develop efficient algorithms and tools to operate on results provided by search engines. The overall objective of this research work is to improve the Web page ranking process.

2. LITERATURE REVIEW

Web Mining is an application of data mining techniques to discover knowledge from Web data. 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. G.Poonkuzhali et al. in their paper **“***Detection and Removal of Redundant Web Content through Rectangular and Signed Approach*” 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.[6] Bing Liu et al. in their editorial issue “*Special*

Issue on Web Content Mining “explained basics of Web content mining. Web offers various challenges to data mining because of various characteristics of Web such as heterogeneous, huge and dynamic nature of Web. Various suggested research topics of Web mining in this issue includes Structured Data Extraction, Unstructured Text extraction, Web Information Integration, Mining Web Opinion Sources. [2] Bhanipriya et al. in their paper “*Web Content Mining Tools: a comparative study* “explained Web content mining and its related techniques and also comparison of various tools for Web content mining is included which helps to understand functionality of various available tools for Web content mining. [3] Hillol kargupta et. al in their book *“Data Mining – Next Generation Challenges and Future Directions* ” explained basics of Web mining and also it describes applications related to the Web Technologies as well as upcoming research directions. [7] G.K Gupta et. al in their book *“Data Mining with Case Studies”* explained various case studies related to data mining i.e. association mining, search engine etc. This literature helps to understand the basic searching methodology and its architecture. He also explained the page ranking algorithm used by Google. This book well explained the terminology of Web mining and various factors to determine the precision of research. [5] J.W. Han et. al in their book “*Data Mining: Concepts and Techniques* “explained various basic concepts on data mining and also it includes various area of implementation of data mining techniques. [8]. R.Cooley et. al in their paper “Web *Mining: Information and Pattern discovery on the World Wide Web*” explained taxonomy of Web mining . Some of the approaches for Web content mining are Agent based approach, Database Approach etc. Web Miner, pattern discovery tool automatically discovers association rules and sequential patterns from server access logs. After pattern discovery, pattern analysis is required which may be accomplished by Web Viz. system, OLAP techniques or SQL like query mechanism as proposed by Web Miner. He explained the architecture of Web Usage Mining which can be subdivided into two parts i.e. domain dependent processes which include preprocessing, transaction identification and data integration components while another division is domain independent application.[11] Soumen chakrabarti et al. in their literature “*Mining the Web: discovering knowledge from hypertext data*” suggested procedure of discovering knowledge from hypertext data through basic searching of data, learning models are proposed in the form of supervised and unsupervised learning. [13]

3. RESEARCH PROBLEM

Tremendous growth of Web is usually treated as tremendous growth of information but it is also accompanied with number of difficulties in extracting useful and meaningful knowledge. [5]Unstructured nature of Web is responsible for making the situations even worse due to absence of a concept such as catalogue in Web and therefore User is dependent on search engines and search directories etc but some of the common problems associated with search engines are (i) Abundance or Scarcity Problem: User usually end up with thousands or even more links if the search topic is popular and sometimes not even a single link. (ii) How to judge the reliability of ranking of Web pages provided by search engine [5]

4. RESEARCH METHODOLOGY

This research work addresses both of above mentioned problems by using improved Web dictionary based approach to determine relevancy and hence ranking of a Web page. In this approach a Web dictionary is implemented from a candidate Web page based on minimum and maximum number of characters in each of the word of search string entered by user. After then the count of number of words found in the dictionary is compared with number of words not found in dictionary so as to determine whether to eliminate the Web page from the search result or not and finally correct ranking of web pages in output is determined through previously spend time statistic as spent by all of the previous visitors of the same Web page in response to similar search queries executed in past.

5. WEB DICTIONARY IMPLEMENTATION- AN EXAMPLE

Consider an example of search string “HOTEL TAJ IN MUMBAI”. Sample web dictionary from candidate web page may be constructed as shown below:

Table 1: Sample Web Dictionary.

2 Letter Words	**3 Letter Words**	**4 Letter Words**	**5 Letter Words**	**6 Letter Words**
AN	ANT	AUNT	ALPHA	ATOMIC
AT	AXE	BOMB	BUFFE	BOXERS
CD	BOX	BOMB	CAMPA	COMMON
CM	BOY	DOME	CAMPA	CAFFIN
IN	COW	FISH	DONER	CAFFIN
IN	**TAJ**	GOAT	DONER	MONKEY
OX	**TAJ**	TOAN	FLOAK	**MUMBAI**
TV	**TAJ**	USIT	**HOTEL**	PHOBIA
XP	**TAJ**	VENI	**HOTEL**	PLANET

Here min =2(“IN”) and max =6(“MUMBAI”) , so Web dictionary from a sample Web page may be constructed like Table 1 considering only those words from candidate web page having number of characters between min and max. Here all the matches are represented in BOLD format. So found =7 and nfound = 0, Hence this Web page will be considered relevant as found > nfound. Here stem words like ‘IN’ are ignored in frequency calculation.

6. SYSTEM DESIGN

The proposed system consists of four modules i.e**.** Module 1: Web Dictionary Implementation, Module 2 and Module 3 for relevancy determination and module 4 for overall priority determination as discussed below:

6.1 Module 1 – Web Dictionary

Module 1 splits user search string into various words. It then counts length of each of the word to find MIN and MAX which represent minimum and maximum number of characters among various keywords of search string. It will implement web dictionary from the candidate web page by allowing only those words having length in between to that of MIN and MAX. In our example, MIN =2 and MAX =6 so web dictionary will consist of all those words having length in the range of 2 to 6 characters.

6.2 Module 2 – Relevancy of Web Page Using Content

This module will determine relevancy of web page from its content. It will count frequency of various keywords of the search string to determine the value of FOUND and NFOUND variables where found represent total frequency of all the keywords with in web dictionary and n found determine number of keywords not found in web dictionary. The difference between values of FOUND and NFOUND will determine the relevancy of web page.

6.3 Module 3 – Relevancy of Web Page Using Time Spent Statistic

This module will determine relevancy of web page using previously spent time statistic by retrieving its value from database. It will calculate new value of time statistic by calculating average of previous value and new value.

6.4 Module 4 – Priority Calculation Module

This module will determine priority of web page by first calling MODULE 3 at first and after then MODULE 2 so overall priority is determined by judging candidate web page twice using two different modules.

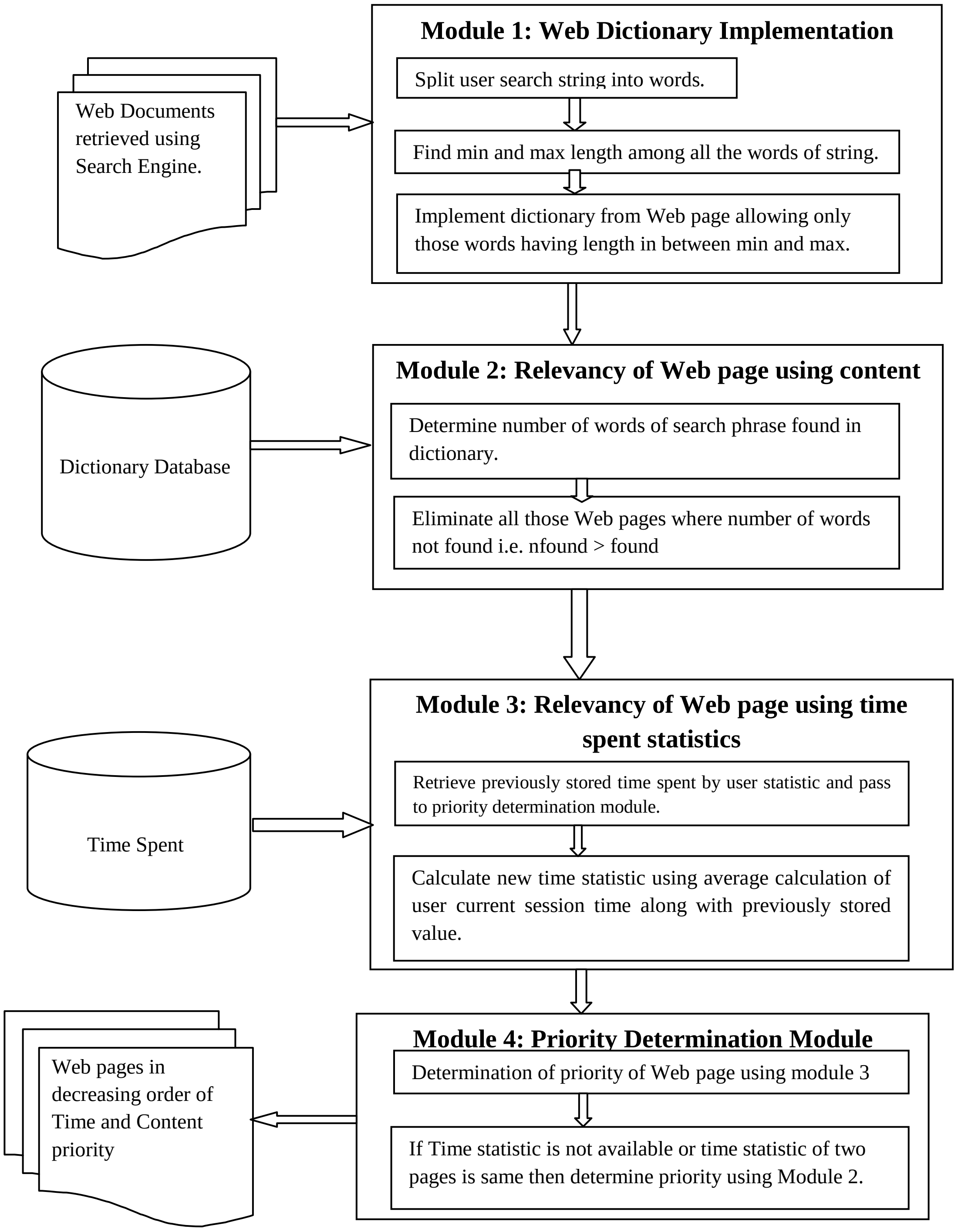
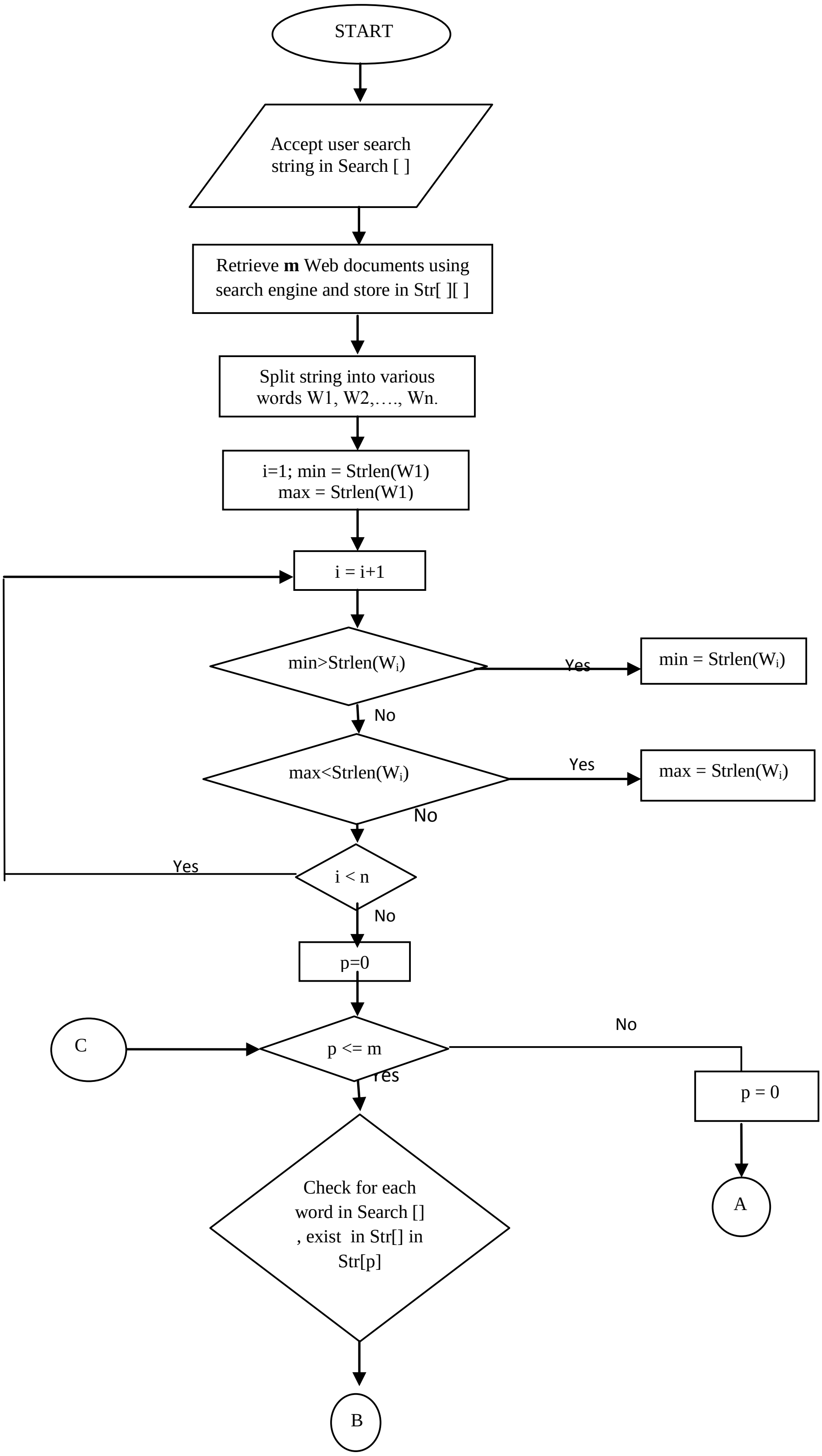

Module 1: Web Dictionary Implementation
Split user search string into words.
Find min and max length among all the words of string.
Implement dictionary from Web page allowing only those words having length in between min and max.
Web Documents retrieved using Search Engine.
Module 2: Relevancy of Web page using content
Dictionary Database
Determine number of words of search phrase found in dictionary.
Eliminate all those Web pages where number of words not found i.e. nfound > found
Module 3: Relevancy of Web page using time spent statistics
Time Spent
Retrieve previously stored time spent by user statistic and pass to priority determination module.
Calculate new time statistic using average calculation of user current session time along with previously stored value.
Module 4: Priority Determination Module
Determination of priority of Web page using module 3
If Time statistic is not available or time statistic of two pages is same then determine priority using Module 2.
Web pages in decreasing order of Time and Content priority

Fig 1.Design of the Proposed System

7. FLOW CHART OF THE PROPOSED SYSTEM


START
Accept user search string in Search [ ]
Retrieve **m** Web documents using search engine and store in Str[ ][ ]
Split string into various words W1, W2,…., Wn.
i=1; min = Strlen(W1) max = Strlen(W1)
i = i+1
min>Strlen(W$_i$)
Yes
min = Strlen(W$_i$)
No
max<Strlen(W$_i$)
Yes
max = Strlen(W$_i$)
No
Yes
i < n
No
p=0
C
p <= m
No
Yes
p = 0
Check for each word in Search [] , exist in Str[] in Str[p]
A
B

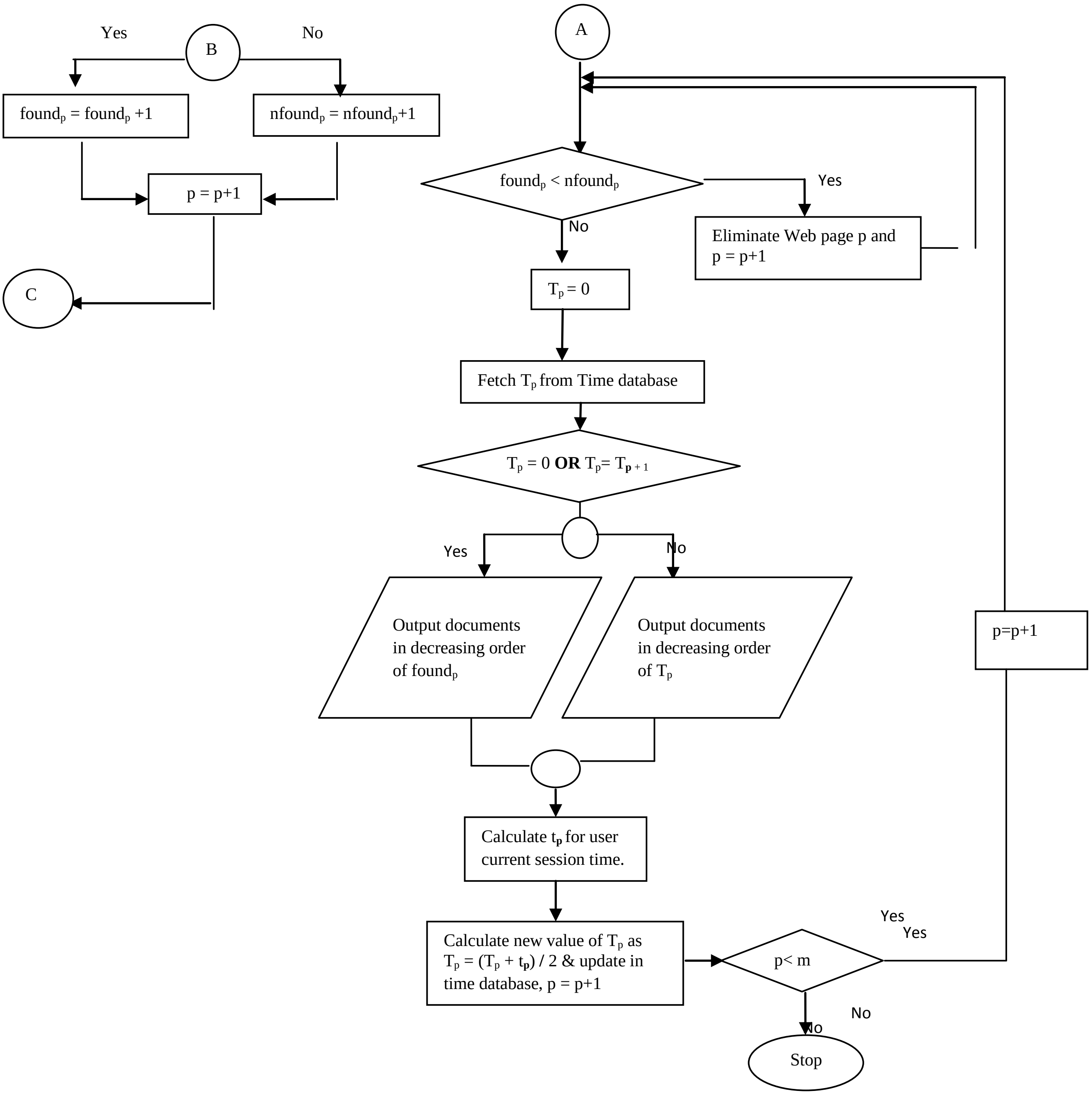

Yes
B
No
$found_p = found_p + 1$
$nfound_p = nfound_p + 1$
p = p+1
C
A
$found_p < nfound_p$
Yes
No
Eliminate Web page p and p = p+1
$T_p = 0$
Fetch T_p from Time database
$T_p = 0$ **OR** $T_p = T_{p+1}$
Yes
No
Output documents in decreasing order of $found_p$
Output documents in decreasing order of T_p
p=p+1
Calculate t_p for user current session time.
Calculate new value of T_p as $T_p = (T_p + t_p) / 2$ & update in time database, p = p+1
Yes
Yes
p< m
No
No
Stop

Fig 2. Flowchart of the Proposed System

8. WEB PAGE RANK DETERMINATION ALGORITHM

Input: User Search Phrase/Query.
Method: Improved Web Dictionary Approach.
Output: Web Pages in ordered priority in terms of contents and average time spent by previous visitors.

Nomenclature

WI – Word in Search phrase.

Dp – Web Document to be scanned.

WDp – Web Dictionary corresponding to Pth Web document.

DW – Document word.

Tp - Average Time spent by previous visitors.

Step 1: Accept search string from user.

Step 2: Search the Web documents (say m in number) using search engine.

Step 3: Split the string into various words W_1, W_2, ……..,$W_{n.}$

Step 4: Determine the minimum and maximum length among the various words of search phrase

```
 min := $Strlen(W_1)$, max := $Strlen(W_1)$
for i = 2 to n do
  if min>$Strlen( W_i)$ then
        min: = $Strlen(W_i)$
  if max<$Strlen (W_i)$ then
max: = $Strlen(W_i)$
```

Step 5: Initialize Ti for each document as 0.

Step 6: Search the time database of tool using keywords entered by user and search for the same documents as given by search engine in previous step to retrieve T_i.

Step 7: Preprocess each Web document D_j in dictionary form WD_j allowing only those words DW_k from D_j which satisfies the following condition min >= Strlen (DW_k) <= max.

Step 8: for p=1 to m do

```
Intialize $found_p$: =0 and $nfound_p$: =0
  if $W_p$ found in $WD_p$ then
        $found_p$: = $found_p$ +1
  else $nfound_p$: = $nfound_p$ + 1
```

Step 9: Eliminate all Web pages where $nfound_p > found_p$.

Step 10: Output remaining Web pages in the decreasing order of T_p and if T_p=0 or $T_p = T_{p+1}$ then in decreasing order of foundp.

Step 11: On start of user session, determine t_p which is session duration of current page and determine new value of T_p as follows:

```
 if $T_p$ =0 then $T_p= t_p$
else $T_p = (T_p + t_p)/2$
```

Step 12: Update the time database of tool with keywords, page address and T_p.

Step 13: Display all the retrieved web pages in decreasing order of rank obtained in previous steps.

9. OBSERVATIONS

As shown below the output of a popular search engine is giving wrong ordered priority i.e. relevant page is listed below comparatively less relevant page. One of the possible reason is the impact of Search Engine Optimization Techniques (SEO) used by various businesses in order to improve the traffic by listing candidate site among Top listed links in Search Engine Output.

Table 2: Sample Output of a Popular Search Engine for a User Specified Search String.

Priority	Link	Found	Nfound	Time Statistic (Second)	Ordered Priority
1	Link 1	140	5	N/A	Wrong
2	Link 2	5	12	N/A	Wrong
3	Link 3	8	23	N/A	Wrong
4	Link 4	500	0	N/A	Wrong

In Table 3 below Link 4 is listed on the top because of User spent average time as well as count of found and however where time statistic is not available(N/A), algorithm is listing such link below other links where time statistic is available.

Table 3: Sample Output of a Tool Using Priority Determination Algorithm.

Priority	Link	Found	Nfound	Time Statistic (Second)	Ordered Priority
1	Link 4	500	0	600	Correct
2	Link 1	140	5	150	Correct
3	Link 2	5	12	5	Correct
4	Link 3	8	23	N/A	Correct

10. CONCLUSION AND FUTURE WORK

Pattern Matching is one among the most growing research areas in Web Mining domain. The explosive growth of Web necessitates improving time and spacing complexity. In this paper we introduced Web dictionary based new algorithm to improve the page ranking process which can work on the output of a search engine. This algorithm requires fetching data from each of the Web page listed intermediately by search engine. However, sometimes due to inherent security implementation of Web servers, Web data is not frequently available to mining algorithms. Also, deep Web which is not directly accessible to search engine and is available only through user query interfaces is also required to be accessed [12]. Therefore future work aims to handle Security and Deep Web issues so as to rank all such Web pages as well.

11. REFERENCES

[1] A. Mendez-Torreblanca, M.Monte." *A Trend Discovery for Dynamic Web content Mining*", IEEE, Intelligence system, Vol. 14, pages 20-22, 2002.

[2] R.Khanchana and Dr. M. Punithavalli *" A Web Usage Mining Approach Based On New Technique in Web Path Recommendation Systems" , International Journal of Engineering Research and Technology, Vol. 2, January 2013*

[3] Bing Liu, Kevin Chen- Chuan Chang, Editorial issue :" *Special Issue on Web Content Mining",* SIGKDD Explorations, Volume 6, Issue 2.

[4] Bharanipriya & V. Kamakshi Prasad," *Web Content Mining Tools: a comparative study*", International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[5] G.K Gupta," *Introduction to Data mining with Case Studies*", PHI.

[6] 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), 4026-4032, 2010.

[7] Hillol kargupta ,Aunpam Joshi, Krishnamoorthy Sivakumar and Yelena Yesha," *Data Mining – Next Generation Challenges and Future Directions* ",PHI, 2007

[8] J.W.Han, M.Kamber, "*Data Mining: Concepts and Techniques* ", New York Kaufmann publishers 2001.

[9] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar, "*Web Mining-Accomplishments & Future directions*", Department of Computer science.

[10] Shohreh Ajoudanian, and Mohammad Davarpanah Jazi," *Deep Web Content Mining*", World Academy of Science, Engineering and Technology 49 2009.

[11] R.Cooley, B.Mobasher, J. Shrivastava,"*Web mining: information and pattern discovery on the World Wide Web*", Department of computer science, University of Minnesota, USA.

[12] Rajshree Shettar, Dr. Shobha G.," *Survey on Mining in Semi Structured Data*", IJCSNS International Journal of computer science and Network security", Vol.7 No.8. Aug 2007.

[13] Soumen chakrabarti," *Mining the Web: discovering knowledge from hypertext data*", Elsevier.

[14] Chen, M.S., Park, J.S. and Yu, P.S., "*Efficient Data Mining for Path Traversal Patterns", IEEE Transactions on Knowledge and Data Engineering*, March/April, 1998,pp 209-221.

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

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