ENHANCING ACCESS OF ARCHIVES AND RANKING IN WEBSEARCH

Archana Shree S.¹ and Vigneshwari S.²
¹Department of Computer Science and Engineering, Sathyabama University, Tamilnadu, India
²Faculty of Computing, Sathyabama University, Tamilnadu, India
E-Mail: archanashreebe@gmail.com

ABSTRACT

In recent days, web searching and security of the archives plays most incredible progress. The enduring research prototypes many web search show the result by searching the relevant data alone. Due to the mere relevancy search, the users may loss some useful data which are not included in the search result. Moreover, it also may consume more time by searching the data sequentially. To overcome these challenges stemming process is united with the existing model for searching the both labelled and unlabelled documents. Furthermore, User Based Advertisement (UBA) is included with the proposed search engine to display the advertisements based on search. To improve the ranking system, Time Stamp Based Analysis (TSBA) is incorporated in the process for easy search of users. With these improvements downloading one's file that are uploaded in the web search become quite easy. But security is a major concern sending a request for downloading a data. In order to overcome this difficulty, Email OTP Alert (EOTPA) is provided in the proposed model to increase the security in web search.

Keywords: web searching, labelled, unlabelled, stemming, data, EOTPA, security.

INTRODUCTION

In prevailing system they used only LAMIS which demonstrates only that appropriate data. To provide best results, stemming model is used to improve the examining development which includes all the labelled and unlabelled data. With this another three processes has been included, in the proposed model the ranking process will be done based on three processes. They are,

1. User query based Advertisement.
2. Time Stamp based Analysis.
3. Loutish word removing.

In the first process, when the user searches the query the related advertisement will be displayed on the web browser. In the second process, it is used to trace the session (i.e. time stamp of a particular user. And in the third process while post the comments in the link any loutish words are used it will give the alert to remove the word. And also with this based on Time stamp based Analysis is used to rank the searching query links by including that how much time the link is used by the user and it is used to improve the examining development as it recycled to reduce the time consumption. Count to this, downloading the archives is easy to users, there is no security provided for the uploading the archives in web search. To overcome these challenges email one time password alert in created, if anyone tries to download thearchives at this time (EOTPA) send the alert message to the concern owner.

LITERATURE SURVEY

Monisha et al [1] says the usage of lamis algorithm they improved the percentage of accuracy and relevance is up to 133% to 232% precision and 0.5 and 1 in recall but there is no discussion on security.

Merlin et al [2] they improved the searching of the keywords by count of number of visitors to the page.

Michael et al [3] says that, they improve the security and privacy risks, which threaten the well-being of OSN common users. And also existing solutions can provide better protection, security, and privacy. But they failed to provide security for the images that they uploaded in web sites.

Hung et al [4] entropy based search that based on LAMIS algorithm by removing the redundant data and the precision percentage is up to 122 to 257 and recall is up to 0.5 to 1. The precision and the recall of Info Discoverer are greater than 0.956 but the drawback is they didn’t mention about any time based algorithm and security.

Vigneshwari et al [5] in novel based stemming process had done for the best improvement in the search on user profiling Ontologies and by this web pages are personalised which are similar in temperament to the user queries.

Nithya et al [6] done novel research in web based mining, they have removed the noise and searching is easier but disadvantage is they don’t have any secure while searching or accessing any documents.

Alper et al [7] proposed the reduction primitives and other encoded algorithms. They had scientific analysis for big data analysis. If modification has done in work like reports, data tools and methods the entire workflow have to be change.

Azad et al [8] explains about the accuracy and relevancy of details extracted from web by semantic synaptic approach. But it is not used for particular link it is used for entire website.

Chaudhuri et al [9] discussed about the complication occurs in SQL database satisfies when so many tuples are provided in queries. So they proposed different ranking approach to improve the quality and performance of the system.

Muthusamy et al [10] says multiple consumers may search for different data are stored in databases in cloud platform shared by many users so the data may not
be secured there may be some clash between the data. They had done a work on providing the data to the particular customer who searches it.

Gowri et al [11] utilize the stemming and tokenization algorithms and hybrid algorithms to improve their search on keywords and easy to implement in java for better viewing. But there may be some loss of data.

Sathyai et al [12] describes that to improve the search engine surf by frequency ranking to mine the web content. Accurateess of the documents can be increased. But reliability of data is not providing in it.

Ayman et al [13] includes the data mining and IR techniques for web document outliers. And they improve the enactment by including n-grams.

Karishma [14] says that without the help of semantic web, exact data extraction is impossible. Every technique has some limitation combination of all techniques may provide some solution.

Latha [15] based on some survey work they provide that none of the search engines gives us the perfect solutions as we search and fails to improve in their updation.

Anjali [16] given the comparative study about the stemming algorithm with lemmatization this stemming has some goals that is to minimize the inflectional forms and derivationally related forms with common forms. And also this stemming is the pre-process of text mining.

Joel et al [17] in the existing system because of numerous publications there are lack of standardizations and also it resulted in confusion of the results then in proposed model they had included the empirical model to increase the performance of the searching of keywords.

Joel et al [18] has added that in their proposed work memory consumption precludes the searching process and also the relationship between execution time and various evaluations.

David et al [19] presented a study about the stemming specifically in English by modifying the approaches.

Dugyu et al [20] has given the comparison of both keyword search engine and semantic search engine with empirical evaluation. In that yahoo shows best result in precision and Google shows best in recall state and finally given the result that semantic search performance is low for both the engines.

Lavanya et al [21] given that the application of tools in the internet by developing the Ajax(more dynamic and interactive) works and programming languages like ASP, Asp Dot in windows os is popular.

Myint et al [22] discussed about the searching of keywords generating and evaluating candidate networks. In this performance evaluation of the proposed algorithm has increased by IMDP and DBLP compared to existing algorithm.

Sudeepthi et al [23] have given the survey details about the semantic algorithm. The prominent part shows how the semantic search engines differ from the traditional schemes then the results will be done.

Rui et al [24] given that single sign on (sso) schemes have some of technical challenges wit lack of access and complexity to browser. And in their proposed system they had included the field study they added the 8 serious logic flaws in ID providers. In this study they proved that sso is insecure.

Pande et al [25] has improved the drawbacks in linguistic stemmer. While compared to this linguistic stemming, n-gram shows better analysis in any language.

Hangiang et al [26] here they explained that there this no instinctive system to filter the data in interacting site.Vigneshwari et al [27] had added that multiple ontology’s various concepts and associative merging and finding the relevant words. And novel framework is used for security of document retrieval based on ontology mining.

Vigneshwari et al [28] had given about the two different ontology’s are Word Net and SWETO. And also it includes that Semantic annotation based on RMS and hashing in cross ontology’s using Rabin Karp fingerprinting algorithm and with different datasets.

Vigneshwari et al [29] included thatSEFOS (semantic enriched fuzzy based Ontological integration system. is used to reduce the searching time by the comparative search and precision rate has improved while query search.

Vigneshwari et al [30] in extraction of multiple Ontologies is based on the concept relationships tries to explain the efficient information search in web by using the Ontology-based mining.

Coffman et al [31] says that while they used relational algorithm, they doesn't have the run time performance level as expected.

Vigneshwari et al [32] shows that the better knowledge about the different Ontologies or domains for that cross ontology mining method has been implemented.

MATERIALS AND METHODS

In existing system LAMIS is used, in that some data may be missed for that by adding stemming algorithm unlabelled data can also be retrieved. In this paper, ranking method is included for the searching process. By the normal search the data which are uploading in the database regarding to the keyword will exist. For example In this if the user searches any keywords, based on that the advertisement will appear on the screen including that in existing system the ranking is proceeded by using the number of visitors to the link but this may not reveal the perfect rank and it may also take time to analysis which link is best by including the time stamp based analysis the time consumption can be reduced. The proposed system includes:

Stemming

Stemming is the route for reducing modified words to their stem, basic or parent format (i.e.) mostly in a written word format. The progress of stemming is known to be as conflation. These sequencers are commonly stated to as stemming algorithms or stemmers. Methods castoff to find out the root/stem of a word:A stemmer for ENGLISH, for instance, should associate the STRING "rats" (and possibly "ratlike", "ratty" etc.) as based on the
root "rat", and "mining","miners","words as based on "mine". A stemming algorithm reduces the words "wishing", "wished", "wish", and "wisher" to the root word, "wish".

\[
\text{MeanWC} = NW / US
\]

MeanWC = Mean no of words per conflation classes, NW = No of unique words before Stemming, US = No of unique stems after Stemming.

**Figure-1.** Ranking process architecture.

**Offensive word removal**

For scrutinising the data, analysis the adult content in web search. The admin of the group includes the entire offence words while browse in admin page. When any comments are posted by the ‘x’ person. The tweet posted by the ‘x’ to be viewed ‘y’ that needs the approval of the ‘x’. Once the approval is accepted the ‘y’ can view the post. Before the posts are viewed by the user, it will be compared with the tweets or posts stored in database and tweeted by the user. If any loutish words contained in the post it will not be directly viewed by the user. In that page some popup block will appear which contains the message the tweet contains offensive words.

**Figure-2.** Pop up block with the message the page contains offensive words.

**ALGORITHM**

This algorithm is physically created ontology’s but only for particular search. The originators should be skilled in domain information.

1. **Step 1:** G = set of pages. For each page p in G do
2. **Step 2:** p.auth = 1 // p.auth id the authority score of the page p
3. **Step 3:** p.hub = 1 // p.hub is the score of page p
4. **Step 4:** function Hubs and Authorities (G)
5. **Step 5:** for step from 1 to k do // run the algorithm for k steps
6. **Step 6:** norm = 0
7. **Step 7:** for each page in G do // update all authority values first
8. **Step 8:** p.auth = 0
9. **Step 9:** for each page q in p.incoming.Neighbours do // p.incoming.Neighbours is the set of pages that links to p/
10. **Step 10:** p.auth + = q.hub
11. **Step 11:** norm + = square (p.auth) // calculate the sum of the squared auth values to normalise
12. **Step 12:** norm = sqrt (norm)
13. **Step 13:** for each page p in G do // then update all hub values
14. **Step 14:** p.hub = 0
15. **Step 15:** for each page r in p.outgoing. Neighbours do // outgoing. Neighbours is the set of pages that p links to //
16. **Step 16:** p.hub + = r.auth
17. **Step 17:** norm + = square (p.hub) // calculate the sum of the squared hub values to normalize// norm = sqrt (norm)
18. **Step 18:** p.hub = p.hub / norm // normalize the hub values.

Time stamp based algorithm is used in the existing system. Using the LAMIS which will be used to retrieve the labelled document in the search, in this only normal ranking process is used. But in the proposed model the ranking is done by time stamp and not by using the number of visitors.

**CONCLUSION AND FUTURE WORK**

Searching process can be improved by using the ranking process in that time consumption can be improved and to that advertisements are added to it. In web search, searching of the queries is not available in this proposed model that is included. The main problem of the users is that offensive words usage by x-person and downloading one’s personal archives and to overcome these problems EOTPA and offensive word removal concepts are included. And our proposed system is efficient in searching of the offensive words in the posts and search then it gives alert.
Table-1. Unordered timestamp search.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Files</th>
<th>Unordered timestamp search (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>java</td>
<td>0.38645</td>
</tr>
<tr>
<td>2</td>
<td>Zzz/java</td>
<td>0.311567</td>
</tr>
<tr>
<td>3</td>
<td>jsp</td>
<td>0.831083</td>
</tr>
<tr>
<td>4</td>
<td>oracle</td>
<td>0.7008</td>
</tr>
<tr>
<td>5</td>
<td>.net</td>
<td>8.09857</td>
</tr>
<tr>
<td>6</td>
<td>Installing java</td>
<td>10.0975</td>
</tr>
<tr>
<td>7</td>
<td>thunder</td>
<td>0.1827</td>
</tr>
<tr>
<td>8</td>
<td>A.P.J</td>
<td>0.2209</td>
</tr>
<tr>
<td>9</td>
<td>HTML</td>
<td>15.098</td>
</tr>
<tr>
<td>10</td>
<td>sachin</td>
<td>0.478890</td>
</tr>
<tr>
<td>11</td>
<td>languages</td>
<td>0.670948</td>
</tr>
<tr>
<td>12</td>
<td>comments</td>
<td>0.0378</td>
</tr>
<tr>
<td>13</td>
<td>Atlanta</td>
<td>5.09756</td>
</tr>
<tr>
<td>14</td>
<td>111</td>
<td>0.04435</td>
</tr>
</tbody>
</table>

In Table-1 here this shows the comparison between the normal retrieval of the archives gives the timing of the system as usual it is added by the admin in the database.

Table-2. Ordered time stamp based retrieval.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Files</th>
<th>Ordered timestamp retrieval time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>HTML</td>
<td>15.098</td>
</tr>
<tr>
<td>2</td>
<td>Installing java</td>
<td>10.0975</td>
</tr>
<tr>
<td>3</td>
<td>.net</td>
<td>8.09857</td>
</tr>
<tr>
<td>4</td>
<td>Atlanta</td>
<td>5.09756</td>
</tr>
<tr>
<td>5</td>
<td>jsp</td>
<td>0.831083</td>
</tr>
<tr>
<td>6</td>
<td>oracle</td>
<td>0.7008</td>
</tr>
<tr>
<td>7</td>
<td>languages</td>
<td>0.670948</td>
</tr>
<tr>
<td>8</td>
<td>sachin</td>
<td>0.478890</td>
</tr>
<tr>
<td>9</td>
<td>Zzz/java</td>
<td>0.311567</td>
</tr>
<tr>
<td>10</td>
<td>java</td>
<td>0.38645</td>
</tr>
<tr>
<td>11</td>
<td>A.P.J</td>
<td>0.2209</td>
</tr>
<tr>
<td>12</td>
<td>Thunder</td>
<td>0.1827</td>
</tr>
<tr>
<td>13</td>
<td>111</td>
<td>0.04435</td>
</tr>
<tr>
<td>14</td>
<td>comments</td>
<td>0.0378</td>
</tr>
</tbody>
</table>

In Table-2 it shows that in the proposed process by using the timestamp the files are arranged in the webpage.
Figure-6. OTP page.

While downloading the file, OTP will be sent to the registered user's email. In the proposed process, time stamp is used to rank the files included in the data by using the time taken by the user to visit the page. With this, files will be displayed according to the time.

Figure-7. EOTPA.

Figure-8 shows the Email OTP that has been approved and sent by the person who uploads the file while that is needed to be viewed by the other registered user that request had sent by the member who has registered in the page.

Table-3. LAMIS search vs. time stamp search for ebay dataset.

<table>
<thead>
<tr>
<th>No. of documents retrieved</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>134</td>
<td>0.6</td>
<td>0.56</td>
</tr>
<tr>
<td>150</td>
<td>0.69</td>
<td>0.74</td>
</tr>
</tbody>
</table>

REFERENCES


[27] Hanqiang Cheng, Xinyu Xing, Xue Liu, and Qin Lv. 2015. ISC: An Iterative Social Based Classifier for Adult Account Detection on Twitter. IEEE Transactions on Knowledge and Data Engineering. 27(4).


[31] Vigneshwari.s and Aramudhan. M. M. 2015. Web information extraction on multiple Ontologies based on concept relationships upon training the user profiles. Proceedings of Artificial Intelligence and Evolutionary Algorithms in Engineering Systems, the International Conference on Artificial Intelligence and
Evolutionary Algorithms in Engineering Systems

ICAEES.
