
# RELEVANT INFORMATION SEARCH THROUGH INTERNET

Ranganathan Hariharan

*Gojan School of Business and Technology, Chennai, India*

*ranlal@yahoo.com*

***ABSTRACT***

*There is exponential growth of information taking place world over. Keeping pace with the growth is the people's quest for more knowledge. Nowadays, even school pupils search the Internet for new information. While searching the net, there are so many links presented. If the person seeking the knowledge is not aware of the basics about the topic of search, the searcher is confused about choosing the correct one from the multitude of information presented. If the person is aware of some basics, the search query can be modified with advanced options available with the search engine. The author proposes a unique search method in which the searcher can decide on the parameters of search and set a relevance factor so that a filter can filter out irrelevant / repetitive information. Key words: Information search, Relevance factor, Web search*

## INTRODUCTION

Every human undergoes a lot of experiences regularly. Every human tries to acquire knowledge through majority of the experiences. Such increase in knowledge is stored in the minds of the persons who have gone through the experiences. Many times, the persons want to share the knowledge with as many people as possible. Such sharing of knowledge has resulted in exponential increase of information world over. Such a phenomenal growth of information is closely followed by the quest of people to to be up to date on the latest knowledge. One finds that the quest for knowledge is not restricted by age. Even the school going children nowadays are accessing the Internet and are seeking knowledge through many search engines such as Google, Yahoo etc.

With vast knowledge being available and with the mobile technology assisting the access of the knowledge from anywhere in the world at anytime, the Internet is busy with the transfer of information at one end of the world to the other especially with the use of varied mobile devices such as Smartphone, other mobile hand-held devices and Laptops etc with mobile modems etc. If one considers the period between 2000 and 2014, it is said that the Internet users have grown 676% (Internet World Stats, 2014). Such exponential growth of Internet traffic has been spearheaded by the ease with which anyone can access the Internet and search without explicit knowledge in programming etc. Many of the access technologies are integrated in various mobile apps that do not expect special knowledge on operations.

With the increase in Internet traffic, one finds that the majority of the use is in searching for information. This is mainly because of the fact that the searchers intuitively feel that the information gives them the competitive edge as suggested by Gary Marchionini (Marchionini, 1995). Everyone is under compulsion to keep abreast of the latest developments as compared to the others in terms of knowledge to continue to be relevant in business, work etc. Internet offers the necessary wherewithal to be abreast on any topic. It is also found that Google accounts for 90% world search traffic in the Internet. It is also said that Google accounts for 3.5 billion queries in a day. As per information provided by Google, 16% to 20% of Google queries are not queried earlier. Google, in turn, has its operations programmed in such a way as to search the other resources and provide the responses to the searcher. It can easily be understood that the volume of data handled is huge. When there is such a vast traffic of personnel occupying cyber space, natural event to follow is that many business houses are interested in displaying the advertisements about their wares in that space. These business houses view Internet as a new avenue for advertising their goods and services and to influence the customer sentiments and behavior. It is not uncommon to see that one's search is taken over by many advertisements.

The search engines search different data resources to bring the results for each query. The data stored in different resources follow their own individual formats (Crowsey et al, 2007). So for processing each query, the search engine has to search different resources and formulate the results in a standard format before the results are displayed. The complete set of operations consisting of gathering information from documents, hyper links and other resources available in the Internet and to extract knowledge / intelligence from these data is known as Web Mining. So, Web mining is carried out using the steps of finding the resources, selection and pre-processing of the information from the resources, analysis and generalization of data to get the information in a way the searcher wants. Text mining is a sub set of web mining. Text mining is a process of analyzing the textual information available from the web and presenting in format as specified by the user, sometimes in Natural Language also (Shi and Kong, 2009). As discussed, the volume of data to be searched is huge and the criteria and formats of search are also many in number. So, text mining is a mammoth task. For giving meaningful results after processing such voluminous data involves the process of structuring the text, identifying patterns from the data and evaluation and interpretation of the data received.

The data mining algorithm involved in text mining has to carry out the process of bringing out the relevance, novelty and interestingness of the output and to present the same interestingly to the searcher. Text mining tasks are text categorization, text clustering, concept / entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (*i.e.*, learning relations between named entities).

In Gary Marchionini's opinion, there cannot be a meaningful search for information without the intelligent and attentive involvement of the searcher. In the case of text mining, web search is th primary and important step. Google is the most widely used search engine. However, the users who use Google have different goals when they use Google for searching. Large entities of users such as corporate organizations, Government agencies etc are using the web search to decipher intelligence buried in layers and layers of voluminous data. The data to be mined is from emails, customer complaints, survey forms, blogs, internal / external reports, voice messages, text messages etc. Buried in such voluminous data is the intelligence required by the searcher. If someone sets to manually sift through the data, it will be practically impossible to make any meaningful information from such data. Normally mining of this data is carried out to get an insight in to Market Research or Customer Relationship Management or Customer Complaints Redress or Sentiment Analysis or Product Promotion etc.

Let us consider an individual searcher who wants information on a specific topic. There will normally be no commercial motive behind such search except that the search will be a general purpose one. That is, the person may not be interested in selling some goods or services after gaining the knowledge through the search. Such searchers also make a considerable part of the all searchers. The basic requirement of such a person will be to get information on a topic or an object or a concept for better insight or to develop a solution for a problem or just to know the state of the art. Such a searcher will wish that the information be available in its natural format without any further formatting by the search engine. If mining of this information or formatting of this information is required, the searcher would want the operation to be completed by the searcher himself / herself.

A problem seen during the search operation is that the results displayed are interspersed with advertisements of some kind or the other promoting some product related to the topic of search. For example, if the search is about "Colombo", the results displayed will contain, among other results, advertisements pertaining to many hotels in Colombo. Definitely, the searcher may not be intended to visit Colombo in the near future. For such a searcher, the interspersed advertisement will act as a deterrent to use the information.

Another problem during the search operation is that the results displayed may not always be relevant to the topic of search (O'Connor, 2006, Zaragoza and Najork, 2009). For an individual as a searcher, it calls for reasoning by the individual to decide more relevant results from the ones displayed and then to go ahead with further search. The person has an option of using advanced options such as AND and OR of different phrases used during the search. Many times, it is found that the searcher is searching in pursuance of knowledge and is presented with results without relation to the topic searched. That is, the searcher is either loaded with more information to handle or scant or no information at all. Sometimes, a simple query throws out enormous amount of information running into thousands of pages of information. This leads to a situation where a person who is searching to gain some basic knowledge is saddled with the responsibility of deciding from the display which are the relevant ones and which are not. This type of situation gets further complicated when the searcher uses more than one word. The searcher has the option to use advanced search options such as AND / OR and narrow down the possible results to a manageable number. However, to decide on which are the most relevant results for the search is decided by the searcher. Many times, the searcher is expected to decide between 'AND' and 'OR' options and the searcher may not be able to make the decision as to which one is better

suited for that particular search. This situation is paradoxical as the searcher is in search of the knowledge and the search engine expects the searcher to be having the knowledge and using it at an advanced level.

Coupled with all the problems outlined is the fact that an average searcher does not go beyond first page of the results. It is said that 75% of the searchers do not go over to even second page of the results displayed. If the relevant information is displayed in the fifth page, the probability of the searcher seeing the result at the fifth page is very close to zero. There are organizations and software packages which influence the results displayed and the order of the display. Organizations would want their information displayed at first page in the beginning itself. This is considered as a form of advertisement for their products or services. The algorithms for such programs use the behavioral pattern of the searcher and try to manipulate the information displayed at first to suit the tastes of the searcher. Owing to such approach, it is not uncommon to see that the results displayed for the same query are not the same when two different searchers submit the query from two different computers / devices.

## LITERATURE REVIEW

A crucial issue such as this has attracted the attention of many software developers and the corporate alike. As indicated earlier, corporate see this as an ocean of opportunity to advertise their goods and services and the software developers are ready to respond to the requirements of the corporate. There are number of software packages available catering to different needs of the corporate houses. These packages access the data from different resources on the web in regular formatted form or otherwise and derive intelligence from the data before they are displayed as results. For example, Lexalytics Text Analytics is a package that offers systems for market research, social media monitoring and sentiment analysis, survey analysis / voice of customer, enterprise research and public policy. Today many such packages are available for web text mining to cater to different needs of different customers. The stress in all these packages is mainly to use the rich web data for mining and carry out market analysis, sentiment analysis, and behavior analysis and to provide information to the organizations / Governments. They can use the results of analysis in a way that suits them best. However, armed with various analyses, when anyone searches the web, these software packages take over the search operation and provide the results that will be more applicable to the 'profile' of the searcher. These are not meant to help an individual searcher with some quest for knowledge or with some urgency to know about some topic.

When a query is initiated in the search engine, the results displayed contain the same information repeated many times either from the same or similar websites. Sometimes the information displayed will also not relevant to the search. It will become a mammoth task for the searcher to sift the required information from the multitude of data presented. Figure 1 explains the point with the search about "Rajdhani Express" in one of our computers. The figure is a screen shot for the results of the search.

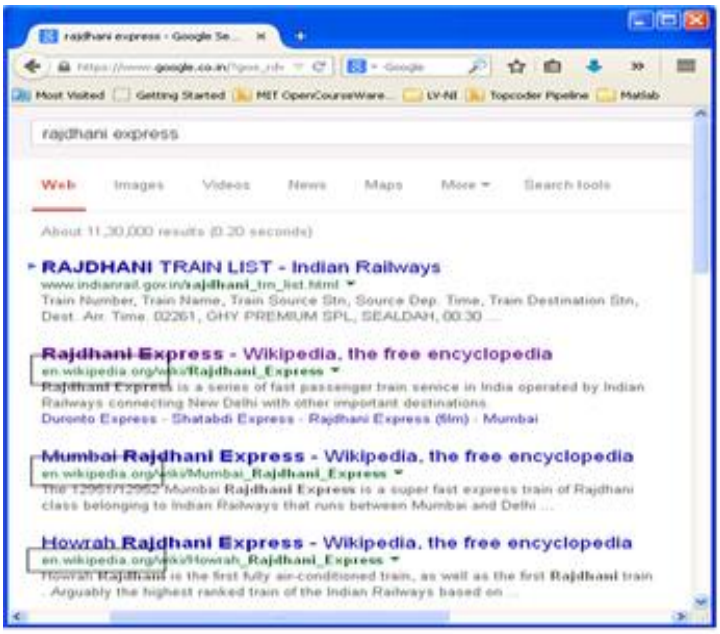


Figure 1: Screen shot of results for the query "Rajdhani Express"

It can be seen from the figure that the second, third and fourth results are the same and from the same link. It can

also be seen from some searches that the results displayed are not relevant to topic of search.

There are many definitions of relevance by many researchers. The relevance is defined for text mining (from a published book such Reuters Corpus volume 1 – RCV1) and also web text mining (from the mining of data from many resources in the Internet). Elastic search considers the number of times the term appears in the returned result, the length of the query and how often the term appears in index to be basis on which the relevance is to be defined. However, Chung and Lui feel that there is a need for structured extraction of results from the web (Chung and Lui, 2001). Ben Carterette and Rosie Jones (Carterette and Jones, 2008) define a model to link clicks received by web search engines to predict document relevance. Fillip Radlinski and Thorsten Joachims (Radlinski and Joachims, 2005) propose a method to link click through data to rank retrieval functions. Discounted Cumulative Gain is considered to be a measure of information retrieval quality. The twin drawbacks of being additive in nature and independence assumption (Chapelle et al, 2009) are the drawbacks in this method. Roa – Velverde and Angel – Sicilia (Valverde and Sicilia, 2014) have compared all web information extraction systems. This can be considered to be starting point for empirical evaluation of ranking of web data.

In considering the search from published book such as RCV1, the definition of relevance during a text mining or information retrieval undergoes a change. Y Li A Algarni and N Zhong (Li et al, 2010) and N Zhong, Y Li and S Wu (Zhong et al, 2012) have considered patterns in text documents as high level features to weigh the effect of them on low level features based on terms or phrases. They report improved performance as compared to other information retrieval algorithms. RCV1 is a resource that is used in many text mining approaches. The extent of volume is clearly defined for RCV1. The analysis of relevance in text mining is defined in many literatures (Dumais, 1991, Lewis, 1992, Sabastini, 2002). Many approaches are based on the concept of 'bag of words', where a number of keywords are used as elements in a vector space. Dumais (Dumais, 1991) specifies a Rocchio classifier where in TF*IDF weighing schemes are proposed for text representation. This bag of words approach is ridden with the problem of having to define a set of features from a large number of alternatives. It is expected that the searcher is having sufficient knowledge to use this technique. Other techniques are also proposed for text mining based on concepts such as Information Gain, Mutual Information, Chi Square, Odds Ratio etc (Lewis, 1992, Sabastini, 2002). X Li and B Liu suggested a pattern based technique for text mining using RCV1 (Li and Liu, 2003). The approach in text mining is different from the approaches that are to be adopted in web text mining. When a searcher searches the web for information, the extent of data is not defined and it is also continuously growing. So a different approach has to be adopted for web text mining applications.

Getting relevant results for web queries is an open research area. A few techniques are reported in web text mining. Aarti Singh proposed an agent based framework for semantic web content mining using clustering techniques (Singh, 2012). M Amarendra and R V S Lalitha describe a method based on some pre-existing knowledge about the database to be mined (Amarendra and Lalitha, 2011). Wen Zhang and Xijin Tang propose a Chinese Web Text Mining process (Zhang and Tang, 2006). However, they have indicated that their work is in the initial stages. C H Lee and H C Yang reported bilingual web text mining for Chinese – English corpora using Self Organizing Map (SOM) neural net that works on the basis of clustering technique (Lee and Yang, 2000). They have also indicated that their work is in initial stage. M Castellano et al present a web text mining method mainly to mine data on job offers (Castellano et al, 2007). As it can be appreciated, there is a crying need for providing a method with which it should be possible to get relevant data when anyone searches the net for any information. R K Srihari et al describe "Info Xtract" as an intermediate engine to extract information from the web (Srihari et al, 2006).

Based on the discussions above, it is clear that there is a requirement for having a mechanism to ensure relevant

results are returned whenever a searcher searches the Internet for any information.

## PROPOSED METHODOLOGY

It is proposed to have an approach that observes the results before they are displayed. The repetitive results with the same web link and the results with lesser relevance are eliminated before the actual display. So, only relevant and non-repetitive results will be displayed. However, many times, the searcher may not be aware of what is relevant and what is not from the plethora of results displayed. The use of advanced options, as seen earlier, may not be a good option either. So, the user is given an option to define a Relevance Factor during the search. To accomplish the same, a filter is designed to operate over the search engine. Similar relevance factors for combinational words can also be used. The filter will carry out the following tasks:

To filter out results from same / similar websites. That is, results from same link addresses will be deleted in the results displayed. This way, the problems as identified in figure 1 can be eliminated.

Users define a Character Relevance Factor (CR factor). The character relevance factor can assume a reasonable value close to 100%. CR factor allows the extent of match between the query word and the result word. For example, if a user sets a 75% relevance factor for a search of the word "Book", the word "Look" will also pass the filter as three out of four letters (75%) match between the query word and result word. CR factor is introduced to cater to the needs of the searchers who may have inadvertently or without knowing full spelling may have entered a query word.

The filter acts as an agent between the searcher and the search engine. It submits the query to the search engine and gets the results from the search engine. After getting the results, the filter applies the logic for display of the results. It filters out the repetitive results and the results outside the CR factor set by the searcher. The steps of tasks carried out by the filter are shown in figure 2.

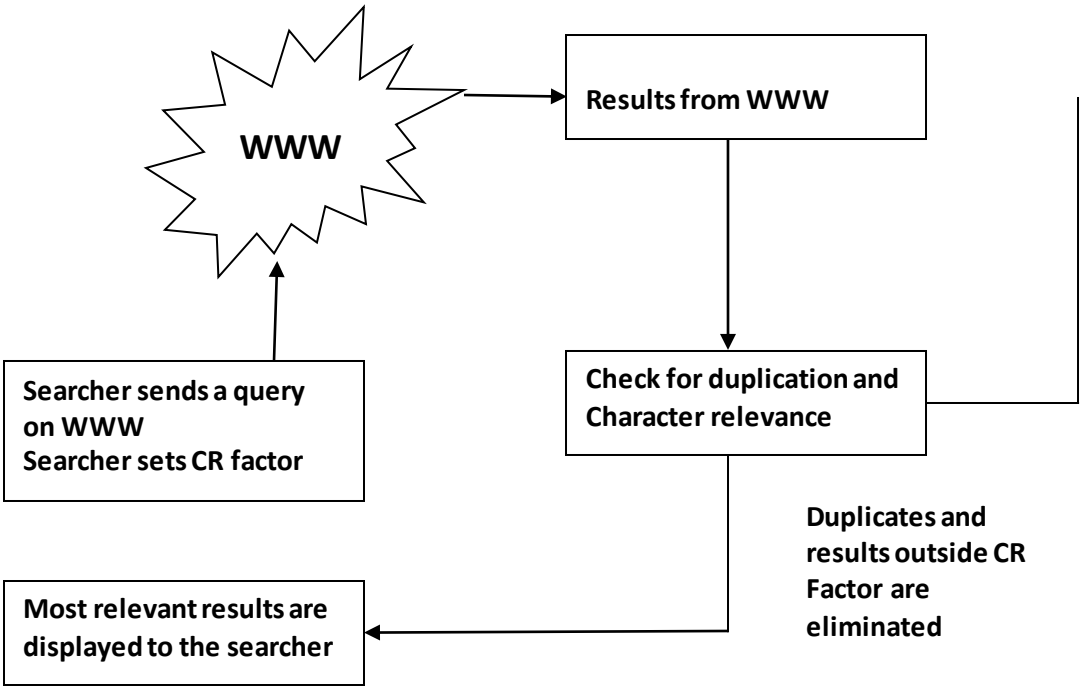

Figure 2: operations of the filter

Based on the flow of task as shown in figure 2, the sequence of steps are as follows:

Define the query

Set Relevance Factor

Initiate a query through Google

Get the output from Google

Check for websites and links for each of the results with the already accepted results

If the links / websites are already in the results, discard that particular result. If not, accept the result.

Go for comparison with the next result to repeat step 5

Once accepted, compare the key words with the query words.

Estimate the percentage of coincidence

For example, if the query is for "pool" and if the result is "tool", then there is 75% relevance

Since the user defines the extent of relevance, this search can get better results as compared to searches with no filter or searches for words within quotes. In the second case, only 100% matches are listed and if the user does not know the exact words, the user is likely to miss relevant results.

The filter is implemented using a program written in Java and operating over Google search. Though Google search is used, the filter can operate over any search engine. Different CR factors are set and tested with different queries. Since the filter operates over the search engine, it does not affect the performance of the search engine.

**RESULTS AND DISCUSSIONS**

It is found that the filter is able to filter out the duplicates in the results and the results that do not meet the CR factor set by the user. Figure 3 shows the results of the query on "Rajdhani Express" with the filter.


Figure 3: Screen shot of Results of query on "Rajdhani Express" after filter

It can be appreciated that the repetitive results and results with lesser relevance are filtered out. The query was carried out with CR factor of 100%. It is said that the searchers do not go beyond the first page of the results to see the information. It can be seen that the repetitive results from Wikipedia are eliminated and many results from the second page (as seen earlier) are shown in first page itself.

Any scientific work's effectiveness must be measured. When the filter is run on Google search engine with the query "Rajdhani Express" set for 100% CR factor, it is found for the first 30 results that there is filtering to the extent of more than 50%. There is no degradation in search timings. If any searcher wants a broader set of results, CR factor can be altered to satisfy the conditions as per the searcher's requirements. However, the effectiveness of the filter has to be measured. Metrics are to be defined for measurement. The definition for relevance is elusive as far as web text mining is considered.

Many techniques are available for checking the relevance of results from a search. But most of the reported work focuses on retrieval of information from books / published work / journals etc. When a person queries the net for getting information, two factors are to be considered for measuring effectiveness. They are the level of relevance of the search results against the junk and the level of satisfaction the searcher gets. Precision and recall are the rates that may be defined to measure the extent of success in a search. Precision can be defined as the ratio of the number of relevant results to the total results. Recall is the fraction of the total relevant results retrieved in a search. Both the measures are difficult to measure accurately and also have limitations. As already discussed, the limitation on the knowledge of the searcher may make it difficult to specify the query to exactly pinpoint the requirement. Use of broad synonyms and not 'relevant' synonyms and use of Boolean operators and not the proximity operators may result in reduced precision and recall. Moreover, the measurement of precision and recall is also difficult. The relevance cannot be clearly defined and so calculating precision from one search is difficult. The relevance is subjective and so clear measurement of precision is elusive. Similarly, it is also not possible to measure recall correctly as the searcher (or any other person) will not be aware of the relevant records available in the whole WWW space searched.

The satisfaction level of getting meaningful result is a measure that eludes definition. The definition of user happiness varies based on the entity that rates the happiness. Search engines rate the happiness based on the return users. E-commerce sites rate the happiness based on their ability to convert the user to become a buyer. Enterprises (Commercial or otherwise) rate the happiness based on the site's ability to provide information to the user within short time. Many of the current search engines concentrate on speed of response (Latency), classifying the user (based on purchase pattern of the user) and method of converting the 'casual browser' to a buyer. There are also some companies who ensure that their products' advertisement is returned in the first page whenever there is a query relating remotely to the attributes of their product. In these discussions, one can visualize that the user's requirement for information (Sometimes urgent) is not one of their concerns at all. The proposed filter approach provides the user with the required tool to define the search and arrive at the result more easily.

There is a requirement to calculate the effectiveness of the filter. For us to estimate the extent to which the filter is effective, we need to have a clear idea about the total number of results without the filter and with the filter. While a Google search on a topic displays the first 10 results in the first page, as indicated earlier, many searchers do not go beyond the first page. To study the effectiveness of the filter, one has to identify the total number of results from the search engine without the filter. After identifying the number, the filter has to be applied and the total results after application of the filter has to be calculated. As shown in figure 1, for a query on "Rajdhani express", the total results obtained is 11,30,000. It has grown to 16,50,000 for the same search recently. It is not possible to use the filter to access the results and apply the filter characteristics. A computer program was written to get the results as a whole with intent of applying the filter to all results. This approach could not be employed as Google does not allow automatic repetitive access of search results. So, each page is searched one at a time and then filtered to get the desired search results.

Moreover, it is found that the results displayed for one user for a query is different from the results displayed for another user for the same query at the same time. This is because the display of results is based on the search history. So the order of display also is not the same for all users. So, only way to find the effectiveness of our filter is to search one page at a time, get the results and to apply the filter to see the number of results filtered out. For checking the effectiveness of our filter in the case of query on "Rajdhani Express", it was needed to repeat this exercise over 1, 13,000 pages. This way, for 'Rajdhani express' query, it is found from the search from our computer that 8 out of 10 results in page 0 are filtered out due to application of CR factor and removal of repeat results. In page 1, 9 out of 10 results are filtered out. The number of results filtered out from

page 0 to 20 varied from 8 to 10. After page 20 up to page 100, majority of the pages saw filtering out of 7 results out of 10. It is found that in the first 100 pages of results, a total of 756 results is filtered out of 1000 results, making an average of 7.56 results being filtered out of every 10 results. The filter is also tried on another query on "Shatabthi Express". The total number of results for the query is 948,000. It is seen that an average of 5.2 results are filtered out for this query.

The approach used in designing the filter compares very well with the number of web text mining approaches detailed earlier. Aarti Singh (Singh, 2012) concludes that the approach proposed by her is only in proposal stage and implementation has not been carried out. Amarendra and Lalitha (Armendra and lalitha, 2011) propose a method that expects the user to have some knowledge about the data to be mined. Zhang and Tang (Zhang and Tang, 2006) propose a method with a lot of promise for web text mining of Chinese with accuracy of 0.8 and conclude that their work is in initial stages, needing improvements. The method described by Lee and Yang (Lee and Yang, 2000) shows some interesting initial results and potential ways for future work. The work of M Castellano et al (Castellano et al, 2007) is specifically for job seekers and as such cannot be extended to search general data from the web. "Info Xtract" described by R Srihari et al (Srihari et al, 2006) is an intermediate tool and it is expected to work with other information retrieval tools. They conclude that it is useful but needs to evolve self learning for adaptation to any domain. As against these, our filter clearly filters out repetitions and results with lesser relevance defined by CR factor.

It is not possible to define the number of data documents available over the Internet matching any query. Internet is a continuously evolving medium and the saturation level has not been reached yet. It is also not clear as to when it will be reached. This is in total contrast with text mining applications where the effectiveness of particular software for a particular aspect can be compared with that of the other software. Moreover, as discussed earlier, the documents available in Internet are unstructured or semi structured.

## CONCLUSION

While organizational needs for getting the information from the search are met by many commercially available software packages as discussed earlier, there is crying need for having relevant results only displayed during a web search for an individual searcher. To this end, a filter is designed that reduces the duplication of results from the same website and the user having the option of defining a CR factor with which it s possible to get the results as per the requirements of the user. This approach combines the speed of search and also human intelligence to get more meaningful results from any search operation over the Internet. In view of the difficulty in defining exactly the measures of Precision and Recall, the proposed approach gives the best rate of 'satisfaction' based on the individual searcher's requirements

## REFERENCES

"Internet World Stats", 2014, http://www.internetworldstats.com/stats.htm (Accessed: December 2014)

Marchionini, G. (1995). Information Seeking In Electronic Environments. Cambridge, U.K.: Cambridge University Press.

Micah J Crowsey, Amanda R Ramstad, David H. Gutierrez, Gregory W. Paladino, and K. P. White," An Evaluation of Unstructured Text Mining Software ", Proceedings of IEEE Systems and Information Engineering Design Symposium, 2007, Charlottes Ville, USA, pp. 1 - 6.

Guoliang Shi, Yanqing Kong, "Advances in Theories and Applications of Text Mining", Proceedings of IEEE International Conference on Information Science and Engineering", 2009, Nanjing, China, pp. 4167 – 4170.

Dennis O'Connor, "Ranking – How are Search Results Listed?", 2006

http://21cif.imsa.edu/tutorials/micro/mm/ranking/index_html ?b_start:int=4 (Accessed: December 2014.

Hugo Zaragoza, Marc Najork, "Web Search Relevance Ranking", Encyclopaedia of Data Base Systems, 1st Edn., Springer, 2009, pp. 3497 – 3501.

Chang, C.-H. and Lui, S.-C., "IEPAD: Information extraction based on pattern discovery", Proceedings of the Tenth International Conference on World Wide Web (WWW), 2001, Hong-Kong, pp. 223-231.

B. Carterette and R. Jones, "Evaluating search engines by modelling the relationship between relevance and Clicks", Proceedings of 20th International Conference on Advances in Neural Information Processing Systems, 2008, New York, USA, pp. 217–224.

Fillip Radlinski and Thorsten Joachims, "Query chains: Learning to rank from implicit feedback", Proceedings of the ACM SIGKDD International Conference On Knowledge Discovery and Data Mining (KDD), 2005, New York, USA, pp. 239 - 248.

O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan, "Expected reciprocal rank for graded relevance," in Proceedings of the 18th ACM conference on Information and knowledge management,2009, New York, USA, pp. 621–630.

Antonio J. Roa-Valverde • Miguel-Angel Sicilia, "A Survey of approaches for ranking on the web of data", J. Information Retrieval, 2014, 17 (4), pp. 295 – 325

Yuefeng Li, A. Algarni, and N. Zhong. Mining positive and negative patterns for relevance feature discovery. In Proceeding of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), pages 753–762, 2010.

http://dx.doi.org/doi:10.1109/TKDE.2010.211

N. Zhong, Y. Li and S. Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 1, (2012), pp. 30-44.

http://dx.doi.org/doi:10.1109/TKDE.2010.211

S T Dumais, "Improving the retrieval of Information from external sources", Behaviour Research Methods, Instruments and Computers, Vol 23, No 2, p.p 229 – 236, 1991

D D Lewis, "Feature selection and feature extraction for text categorization", Proceedings of the Workshop on Speech and Natural Language, p.p. 212 – 217, 1992.

F Sabastini, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, Vol. 34, No 1, pp. 1 – 47, 2002

X Li and B Liu, "Learning to classify texts using positive and unlabelled data", Proceedings of International Joint Conference on Artificial Intelligence (IJCAI –'03), p.p 587 – 594, 2003

Aarti Singh, "Agent based framework for Semantic Web Content Mining", International Journal for Advances in Technology, Vol 3, No 2, pp. 108 – 11, April 2012.

M Amarendra, R V S Lalitha, "Web based Text Mining", Computer Engineering and Intelligent Systems, 2011, available online at www.iiste.org

Wen Zhang and Xijin Tang, "Web Text Mining on XXSC", Proceedings of Knowledge and System Sciences – IKSS2006, pp. 168 – 175, Beijing, Sep 2006.

C H Lee and H C Yang, "Towards Multilingual Information Discovery through a SOM based Text Mining approach", Proceedings of PRICAI Workshop on Text and Web Mining, Melbourne, Australia, pp. 80 – 87, 2000.

M Castellano, G Mastronardi, A Aprile and G Tarricone, "A Web Text Mining Flexible Architecture", International Journal of Computer Science and Engineering, Vol 1, No 4, pp. 252 – 259, 2007

Rohini K Srihari, Wei Li, Thomas Cornell and Cheng Niu, "Info Xtract: A customizable Intermediate level information extraction engine", Journal of Natural Language Engineering, Vol. 14, No. 1, pp. 33 – 69, 2006