Received: April 2, 2017

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Effective Mining Approach to Produce Quality Search Results Using Proposed Approach

Dhivya Elavarasan[1]* **Durai Raj Vincent[2]**

[1]*Computer and Communication Engineering, Vellore Institute of Technology, India*
[2]*School of Information Technology and Engineering, Vellore Institute of Technology, India*
* Corresponding author's Email: pmvincent@vit.ac.in

Abstract: Web mining is an application of data mining techniques to extract and process knowledge from sources such as web documents, hyperlinks, website usage logs etc. Due to the outbreak of extensive information through web resources, mining of semantically relevant data for a search keyword is one of the most intriguing studies to research. This work deals with automatically extracting the information from web documents using web content mining. The extracted data needs to be preprocessed in order to obtain the appropriate data format for further analysis. Generally, the content based search algorithm is used to find the items relevant to the keyword searched resulting in an indexed set of similar results. Further, the clustering of the similar data is done by adopting the quality threshold clustering algorithm assigning a similarity index to each of the result items. For the final list of items obtained, the weighted page ranking algorithm is applied to rank the most frequently searched item in the lists. The proposed work efficiency will be determined by the cluster's quality and the query blocks ranking efficiency. Various metrics like cluster purity, NMI, Rand Index, F-Measure, wPRF are used to evaluate the quality and the ranking efficiency of the search result obtained. Duplicate result sets are handled and are castigated for better unambiguous results. The results obtained is proved to be better and to overrun the existing approaches like QFI and QFJ. The quality of the result set obtained is further evaluated by repeating the process considering only the top n ranked items, shuffling the top items or by randomly selecting the items. Thus, enabling to validate and uphold the results of the proposed work surpassing the existing algorithms.

Keywords: Web content mining, Content-based search, Quality threshold clustering algorithm, Weighted page ranking algorithm, Cluster purity, NMI, Rand-index, F-measure, wPRF, QFI, QFJ.

1. Introduction

Web mining is the process of collecting and integrating information from various sources around the web documents through the conventional data mining techniques. In the proposed work an efficient search mechanism is adopted to obtain the most relevant search results for the given specific keyword. In turn, to achieve the best results for each keyword, a set of results or query blocks are created relevant to the search item. Query blocks can provide significant and appropriate results to a search keyword and further enables us to search in several ways and aspects significantly enhancing the search results. Using this approach initially in an appropriate manner the query blocks along with the actual search results can be unveiled. Through this approach, the users can interpret the significant prospect of a query in the first attempt rather than browsing through several pages. This query blocks also enables to search for any vague and uncertain search words. The query blocks resulting in the structured and organized knowledge can be enhanced for semantic and entity search [1] in addition to conventional web search.

It is evident from the fact that important information about the queries are presented as lists and it is frequently retrieved among the most recurring documents. Thus, this proposed work makes use of an aggregate list of search results for a keyword thus providing enhanced and best search list. The proposed system extracts data from the web containing texts, links and documents. The data extraction process is automated using python and selenium which reduces the time in gathering the data manually. preprocess the data to obtain all significant information. The data is normalized to reduce the data redundancy and improve the data integrity. A similarity index based on the search keyword is calculated for each result to group the most relevant information and to filter out the irrelevant information at the preprocessing phase itself. The content-based similarity approach is used to obtain the exclusive or unique query results. There are possibilities for the lists obtained from different websites to be duplicated as they copy the contents changing only the headings. Considering all these possible negatives, ranking becomes tedious. Hence Content-based similarity approach is adopted to track the exquisite correlation between the search lists obtained. The quality threshold clustering algorithm in the proposed work gives better results against noise and outliers than the other algorithms. The work focuses on enhancing the quality of the clusters and improving the ranking efficiency through various evaluation methods.

The proposed work also ensures that the search criteria is not restricted to a particular domain but is a generic methodology with enhanced search. Rather than using a predefined schema the results are considered from the top search items on the retrieved lists. This in turn, leads to the fact that every search keyword will have different or varying query blocks. The results are evaluated with various evaluation methods and the experiments is repeated with different set of ranked data to determine the best method. Rather than taking only the top N items the results are verified by randomly selected items and evaluating the quality. This is explained in detail in the experimental results section.

The remaining paper is further organized as follows. The section 2 gives the related work details in a nutshell. Following this the proposed work is briefed in section 3. Section 4 gives details about the various evaluation methodologies. Experimental results are showcased in Section 5. The work is concluded in Section 6.

2. Related work

2.1 Developing and suggesting query search results

Systemizing and endorsing a query are one of the well-known methods which enable the users to interpret or express the information they require. Modifying a query is one of the best ways to produce the search results relevant to the user's needs [2]. This method is used to aggregate the information present in the query search. Endorsing or recommending is a process of providing relevant search results to the user requested query [3]. This approach is used to further enhance and expand the search criteria for the keywords. The important objective of formulating various query blocks is for a particular keyword finding syntactic relevant information.

2.2 Abstraction of query based results

Query blocks provide explicit aggregated or summed-up results when compared to the conventional summarization algorithm [4]. The conventional algorithms will provide the summed-up results by taking results of their query building methods like subjective or abstract, the number of documents considered as sources for queries, the relationship between search keywords and aggregated results [5]. The proposed method enables to identify the vital points for a keyword and there will be query blocks created for various aspects to the given search word. Instead of obtaining the query result from conventional methods multiple groups of syntactically related item lists can be obtained.

2.3 Information search based on entities

The process of entity search is in consideration in recent times. The objective of this is to provide information that concentrates on entities [6]. Preparing the search results and optimizing a search item can be entity based. There are certain cases where the entity for searching a keyword are taken from the structure of the web pages [7]. The process of finding the query blocks are different. The query block will provide different types of results in addition to entity related information.

2.4 Mining information based on query blocks

Searching by query blocks is a procedure that enables a user to brief, interpret and handle multi-dimensional data. This technique can be extensively integrated into e-commerce and digital libraries. There are various methods enabling to find query blocks using the search keywords [8-9]. In the proposed method, the query blocks are determined automatically for the open domain search keywords when browsing in a general web search engine. The query blocks are automatically mined from the web search results without any additional domain information which is useful for the users requiring more semantic relevant information.

3. Proposed work

In the proposed work the search results for a given keyword can be enhanced based on aggregating the results from various query blocks. The information is collected in the blocks in the form of a list. The top results are retrieved from each block depending on the similarity index and the final enhanced result is presented to the users. The search results are aggregated as lists because web based essential and important

information is consistently accessible in the form of lists. Most websites furnish the essential information in the form of lists which enables to easily segregate valuable and erroneous results further enabling in enhancing the quality of the query blocks. The system has four phases which involve extracting the data from the web and processing them to obtain the better search results. The various phases involved contributing to the system architecture are:

- Data Extraction
- Data Pre-processing
- Clustering
- Ranking and Evaluation

3.1 Data extraction

Data extraction phase is the process of extracting the data from the web documents based on the keyword given for search criteria. The various unstructured sources include web pages, emails, PDF's etc. There are various methodologies to extract data from the web. Regular expression-based approaches to analyze the data, using web page structure or table-based approach, by means of analytics using text and associating the information with other existing information. In the proposed work the text mining method to scrape the data from the web using python and Selenium web driver is used to obtain the data. The data extraction is automated to fetch the data from the web documents. Two files are passed one containing the keywords whose data that should be parsed from the web and the other containing the site details from which data is scrapped. On passing these two files as input and by using the selenium web driver data is automatically extracted from the site.

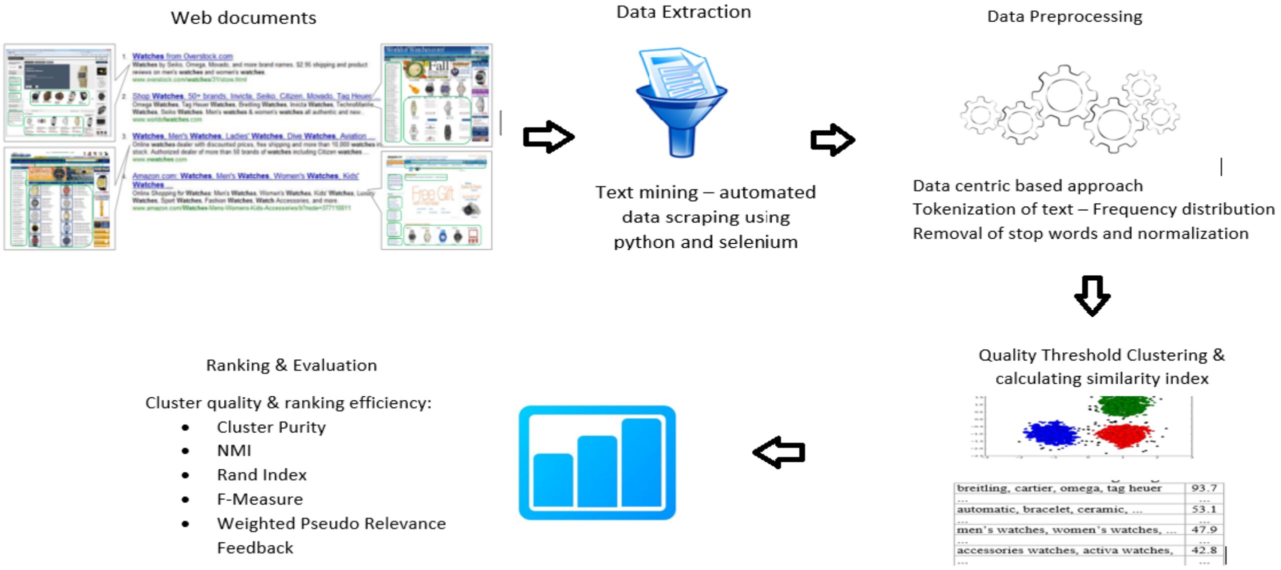
Context based searching is done where the entire pages are scanned and it provides an indexed result based on the text relevance for the search keyword provided. These scraped data consist of details like the keyword, link of the page which is available, the main description of the search result topic, the maximum no of search results etc. From these files, we extract only the relevant information for our work like the main description, maximum searched result, sub data description like only the relevant metadata [10].

3.2 Data preprocessing

The data obtained from the web documents is raw, no quality and accuracy. To structure the data as per the requirement, data preprocessing is done. In the conventional process, the preprocessing first involves removal of duplicates to eliminate redundancy. Data normalization enables to improve the integrity of the data. In the proposed work data is mined in a data-centric approach for better quality [11]. For data cleaning process stop words are identified and removed returning better processed data. Data reduction is achieved by assigning a frequency distribution to the tokenized strings and removing the least frequent or irrelevant data. Now to uniquely identify the processed data a unique id for each of the record is assigned.

3.3 Clustering

Clustering is the practice of converting a collection of hypothetical objects into a collection of relevant objects. Clustering has an advantage of adjusting to the changes in data and generating important features for identifying different groups. Quality Threshold clustering algorithm (QTC) is used to determine the largest cluster for the given data without pre-determining the number of clusters. The termination criteria for the algorithm depends on the cluster diameter. The advantage of QTC over other clustering algorithm is that instead of considering all the data points at a time it considers n randomly selected data points. QTC also gives better results against the introduction of noise and outliers. To reduce the complexity of the QTC algorithm preprocessing of the data is done before the clustering process. Instead of considering the diameter of the cluster the radius from a central point is considered. The neighbor data points are determined from a small set of preferred midpoints thereby reducing the number of calculations per individual clusters. Here a similarity index is calculated for every data. Depending on the similarity index the data is segregated as either exactly matched or partial or not matched data. The quality threshold algorithm works as follows:


Web documents
Data Extraction
Text mining – automated data scraping using python and selenium
Data Preprocessing
Data centric based approach
Tokenization of text – Frequency distribution
Removal of stop words and normalization
Quality Threshold Clustering & calculating similarity index
Ranking & Evaluation
Cluster quality & ranking efficiency:
Cluster Purity
NMI
Rand Index
F-Measure
Weighted Pseudo Relevance Feedback

breitling, cartier, omega, tag heuer	93.7
...	...
automatic, bracelet, ceramic, ...	53.1
...	...
men's watches, women's watches, ...	47.9
...	...
accessories watches, activa watches,	42.8
...	...

Figure.1 Proposed system architecture

- Initialize the threshold distance for clusters and the minimum cluster size.
- Constructing a candidate cluster for every data point by adding the closest point, next closest point and so on, until the distance of the cluster exceeds the threshold.
- The candidate cluster with more points is taken as first true candidate cluster and based on further consideration remove the other points in the cluster.
- Repeating with a reduced set of points resulting in a minimal cluster size until no more clusters are formed.

The advantage of using quality threshold algorithm is all possible clusters are considered. The candidate cluster is generated with respect to every data points in order of size and quality criteria. Once the data is segregated as exact or partial or dissimilar data they are subjected to the next phase ranking and validation.

3.4 Ranking and evaluation

The fourth phase is the ranking and evaluation which enables ranking the processed search results and giving the preference to the data with the highest similarity. In general page ranking algorithm is used to score the links obtained depending on their relevancy. Some of the disadvantages of the page ranking algorithm are that giving more importance to old pages if the new page is not part of the existing website. This may also cause random sinks and dangling links when a network page gets indefinite links or with no outgoing links. To overcome all these drawbacks the proposed work use the weighted page ranking algorithm which is a modification of the page ranking algorithm [12]. This method will assign the highest rank to the most popular pages searched instead of dividing the rank values evenly among various pages. Based on this the top ranked pages for each of the query blocks can be obtained. The evaluation of the proposed work involves two criteria they are cluster quality and effectiveness of the ranking system.

4. Evaluation techniques

The results obtained from the search results aggregated into frequently occurring lists in the query blocks needs to be evaluated to validate the results. The ranking efficiency can be determined using the weighted page ranking algorithm [13] and weighted Pseudo Relevance Feedback (wPRF). The cluster quality is considered and evaluated using various methods like cluster purity, normalized mutual information, rand index and F-Measure.

4.1 Data set used

In order to evaluate the query blocks result there are no any common data set. The dataset is built from scratch. A list of commonly searched keywords are taken and are classified as Userlist. To avoid the intolerance in the query search results another list of frequently search words from a commercial website search engine as Randomlist. Initially, for a given keyword the query blocks are created manually depending on the knowledge on the keywords by considering various resources like

Freebase, Wikipedia or any other sites. A miscellaneous block will be created for each search word collecting the unrelated or bad data. The query blocks which are formed from the userlist and randomlist are now evaluated based on ranking each of the blocks for the given search word in the range of 0 to 2. 0 for bad, 1 for fair and 2 for good respectively. The miscellaneous query blocks are automatically given the rating as 0 and are not subjected to further classification. From Table 1 and Table 2 it is evident that there were nearly 5.1 good and 4.7 fair query blocks when considering the userlist but there are only 3.4 good and 3.3 fair in the random list. There are more query blocks in users list than in the random list because the random list items are randomly sampled from the website search engine.

4.2 Weighted page ranking algorithm

The weighted page ranking algorithm is the modification of page ranking algorithm. The page ranking algorithm will assign ranks to each page depending on the data and links which are incoming and the popularity of the outgoing links. The weighted page ranking algorithm considers the similarity of the query searched results. Thus, in turn assigning different scores for each page rather than distributing the score evenly among the pages. While considering the ranking of the page here we are considering the links available in the page to determine the ranking. The popularity or the reputation of the links relevant to the search keyword is considered. Depending on the relativity or the relevance the ranking is done. In weighted page ranking algorithm, the weights $Wt_{in}(x,y)$ are assigned to the links which are represented as input to a web-based graph, containing the pages to be ranked and to the outgoing links $Wt_{out}(x,y)$.

$Wt_{in}(x,y)$ which is mentioned in Eq.1 is the weight assigned to the incoming link data to a web page where here y represents the page for which the rank is being estimated and *x* denotes the number of pages or links associated with the given page *x*.

$$Wt_{in}(x,y) = \frac{N(y)}{\sum p \in V(x)\, N(x)} \qquad (1)$$

In Eq.(1): *N(y)* is the number of input links for the page y which is considered for ranking, *N(x)* is the number of related pages associated for a given search word for the input link and *V(x)* is the total number of pages to be ranked.

Table 1. Statistics of the dataset created – userlist

Categories	Userlist		
	No. of search keywords (N) = 95		
	Good	Fair	Bad
Search items	25,985	22,741	18,174
Query Blocks	487	450	397
Query Blocks /N	5.13	4.74	4.18
Search items /N	273.53	239.38	191.30

Table 2. Statistics of the dataset created – Randomlist

Categories	Randomlist		
	No.of search keywords (N) = 125		
	Good	Fair	Bad
Search items	17,854	13,972	11,167
Query Blocks	431	417	392
Query Blocks /N	3.45	3.34	3.14
Search items /N	142.83	111.78	89.34

$$Wt_{out}(x,y) = \frac{M(y)}{\sum p \in V(x)\, M(x)} \qquad (2)$$

In Eq.(2): *M(y)* is the number of output links for the page y which is considered for ranking, *M(x)* is the number of related pages associated for a given search word for the output link and *V(x)* is the total number of pages to be ranked.

Equation (3) represents the weighted Page ranking algorithm (WPRA). Here *d* is the damping factor which is usually set to 0.85 where d can be taken as the possibility of the links which are precisely related to the search key and (*1-d)* as the possibility of the links which are ambiguously related to the search key.

$$WPRA = (1-d) + d \sum Wt_{in}(x,y) \text{ x } Wt_{out}(x,y) \qquad (3)$$

In this manner weights are assigned to each page and ranks are allotted based on the weight to each page and the page with the highest weight will be the most relevant to the search keyword.

4.3 Cluster evaluation techniques

The clusters formed with the data must be evaluated to ensure that the cluster contains the most similar data along with them and there are less number of outliers. The main objective of the evaluation method is that the distance between the members of an individual cluster must be more similar than the distance between the adjacent clusters. The various cluster evaluation methods followed in the proposed work are Cluster Purity [14], Rand Index [15], Normalized Mutual Information [16], and F-Measure [17].

4.3.1. Cluster purity

The cluster purity measure is a straightforward and elementary method in assessing a cluster. It is a peripheral cluster assessment criteria which determine the percentage of the data that were correlated precisely. These were rated in the unit range of 0 to 1 indicating no correlation and strongly correlated. To determine the correctness of the measure or assessment criteria firstly a confusion matrix is constructed for the data. The confusion matrix is generally used to determine the performance of the model. We can determine the purity of the cluster using the following formula as in Eq. (4).

$$Purity = \frac{1}{N} \sum Max \; |C_k \; \cap \; O_k| \qquad (4)$$

4.3.2. Normalized mutual information

Mathematically mutual information can be represented as an associated dependence between two members of a dataset. To be more precise mutual information evaluates the amount of information obtained from one data in a dataset through another data. This in turn, is related to the entropy of the data. This is the Normalization of the score obtained from the mutual information for a particular document or a data ranging in a scale value of 0 to 1. Where 0 representing no mutual information and 1 indicating perfect correlation among the data. Mutual Information *I(A,B)* is given by Eq. (5).

$$I(A, B) = \sum_{\mathrm{a \in A}} \sum_{\mathrm{b \in B}} \mathrm{p(a, b)} \log \frac{\mathrm{p(a,b)}}{\mathrm{p(a)p(b)}} \qquad (5)$$

This can be alternatively represented in terms of entropy as in Eq. (6).

$$\mathrm{I(A; B) = H(A, B) - H(A|B) - H(B|A)} \qquad (6)$$

The Normalized Mutual information can be obtained by the normalization of the variables contributing for mutual information.

$$N_{ab} = \frac{I(A;B)}{H(B)} \quad N_{ba} = \frac{I(A;B)}{H(A)} \qquad (7)$$

4.3.3. Rand index and F- measure

Rand Index measure can be used to determine the accuracy of the clusters. Rand Index calculates the percentage of correct decisions. In other words, Rand Index determines the cluster correlation. The Rand Index handles FP and FN with equal weights.

$$\text{Rand Index} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FP+TN+FN}} \qquad (8)$$

The F-Measure enables to determine the accuracy of a method or test considered. The precision and the recall are considered to determine the F-Measure. Precision as mentioned in Eq.9, is defined as the proportion of a number of correct positive results to the total number of positive results in other words positive predictions. The recall represents how many of the positive results were correctly predicted.

$$Precision = \frac{TP}{TP+FP} \qquad Recall = \frac{TP}{TP+FN} \qquad (9)$$

The F-Measure is the harmonic mean which is the average of both the Precision and recall. The F-Measure spans in the scale of 0 to 1. Where 0 indicating no accuracy and 1 as the most precise or accurate. The harmonic means is the convenient way for averaging the proportions.

$$\text{F Measure} = 2 \text{ x } \frac{\text{P x R}}{\text{P+R}} \qquad (10)$$

In addition to the above metrics, the evaluation metric Weighted Pseudo Relevance Feedback (wPRF) is used to determine the ranking efficiency of the clusters. This wPRF considers the various ratings related to each query block. It makes use of the precision, recall and F-Measure. All the above-mentioned metrics are calculated depending on the top results of the search keyword in the query block.

5. Experimental results

In this section, it is focused in mining the user relevant query search results for the given search keywords from the websites. The comparison of the proposed work with various other existing work and the context similarity based approach can be briefly discussed in this section. The following table 2 shows the experimental results of the proposed work considering various ranking and cluster evaluation metrics.

From the Table 3 and Fig 2, it is evident from the fact that the cluster purity is more efficient for the userlist queries than the randomlist queries. They also have better results for RI, NMI and F- Measure for the userlist rather than the randomlist. From this result, it is evident from the fact that the query blocks formed from the basis of the same type of groups are better. Ranking of the list is obtained through the weighted page ranking algorithm. The key points to be considered when ranking the lists in the query blocks are some results may not occur in the list styles or not available in the top search results. It is considered to evaluate depending on the qualified individuals for every query block.

From the Table 3, it is ensured that the weighted page ranking algorithm is used to obtain the better and relevant in the lists which form the query blocks. Thus, it is found that the query blocks generated are essential and appropriate to the users to interpret and recognize their search results.

5.1 Comparing the proposed work with the existing work

The proposed work results are compared with two already existing algorithms Query Faceting Independent (QFI) and Query Faceting Joint (QFJ) [18]. The algorithm QFI forecasts the results by considering whether the given search keyword is present in the screened list and also considering the fact that whether two individual items need to be aggregated in the query block separately. The second algorithm QFJ enables to perform joint intervention by relatively increasing the target value. The search result list which is obtained by the proposed method is compared with the list obtained from QFI and QFJ. The features and standards which are followed by the proposed work are used for comparison in the existing approaches also. The proposed algorithm improves the process by efficient clustering and removing outliers.

The experimental results comparing the proposed work with the existing work are displayed in table 4 and in Fig 3. It is evident from the results that the proposed work outruns the existing algorithms QFJ and QFI in turns of wPRF. It is observed to get persistent results in the dataset containing userlist.

Thus, for the proposed work it is observed that the userlist is providing a much better value of 0.568 than the randomlist dataset value of 0.421 as they are randomly sampled from a website. In order to further evaluate the efficiency of the system, the list of items in each of the query block is subjected to a trial and error method of:

Table 3. Cluster quality and ranking efficiency

Query blocks quality and ranking efficiency		
	userlist	Randomlist
Cluster purity	0.758	0.715
NMI	0.842	0.743
Rand Index	0.755	0.697
F- Measure	0.862	0.792
wPRF	0.743	0.623

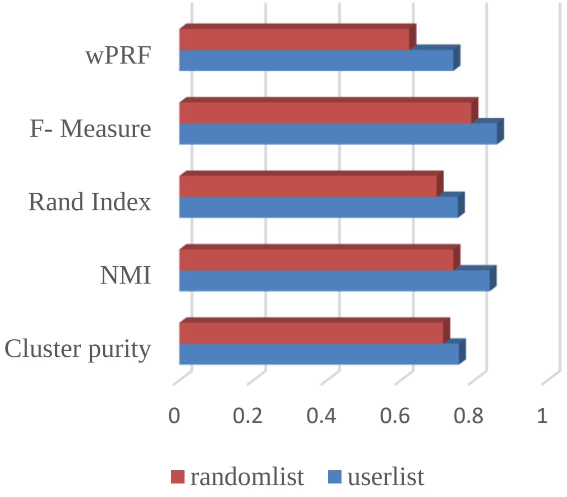


Figure.2 Cluster quality and ranking efficiency

Table 4. Comparison of wPRF values with existing approaches

Comparison with QFI and QFJ wPRF values		
	userlist	Randomlist
Proposed work	0.568	0.421
QFI	0.241	0.354
QFJ	0.397	0.216

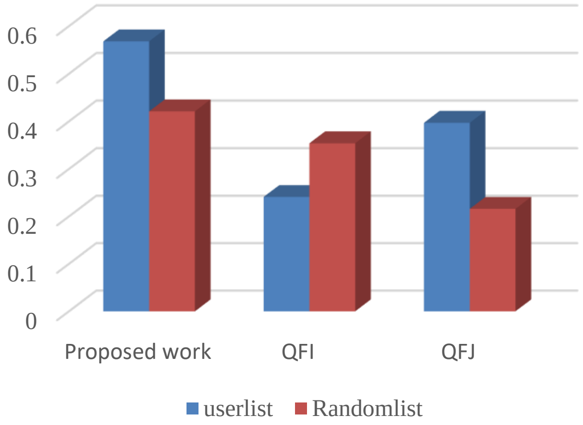


Figure.3 Comparing proposed work with existing work

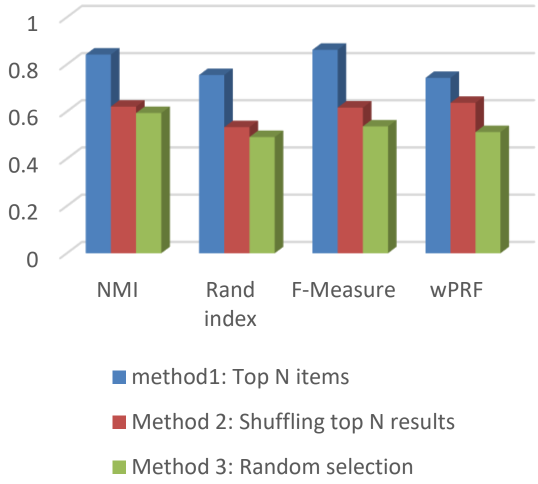


Figure.4 Experimental results with the query blocks quality

(1) Considering only the top N results
(2) Shuffling the top N results randomly
(3) From the query blocks randomly selecting any N items and then using the shuffled items from them.

When considering the results of the above three methods the third method has very less ranking efficiency than the other two methods. The best results can be obtained from the method one which involves considering the top N results. This can be observed in the Fig.4.

Since the third method generates query blocks with very less ranking efficiency it is evident from the fact that the query block contains list items which are less consistent to the given search keyword and also have less efficient items. When looking into the second method of shuffling the top results randomly may result in fetching the item for the query block from less significant documents further leading to the degradation of the quality of the query block. From these experiments, it is evident that the quality of the searched items for the list does affect the quality of the query blocks.

6. Conclusion and future work

In the proposed work, it is focused on extracting efficient query blocks for a given search keyword. An organized and systematic workflow has been practiced in dynamically extracting the query results from the web. Aggregation of relevant and persistent search items from the free text in the web and resulting in top search results for the given search query. Using the modified quality threshold algorithm considering only the radius of the cluster from a central point thus determining the neighborhood points from smaller set reducing the number of iterations to be performed. Two datasets have been formed and used. One containing user defined list of search words and the other randomly samples results from the web. Various measures like the cluster purity, rand index, NMI and wPRF has been used to determine the quality of the cluster and the ranking efficiency of the queries. The proposed work clearly shows the final list of data obtained from the proposed system is found to have better ranking efficiency and cluster quality than the existing algorithms (refer Fig.3). Further the quality of the query blocks obtained is evaluated by repeating the experiments by fetching top N items and random items concluding the results obtained by the top N items are better and efficient (refer Fig.4) concluding the approach provides better results in the first search iteration.

The proposed work can be improved in several ways. The process of finding the query blocks for the given search keyword can be further improved by using any semi-supervised bootstrapping algorithm for list extraction. Explicit wrappers for some websites can be used to obtain highly efficient and qualified list from the websites. This in turn, may show significance in terms of accuracy. The quality of the query blocks can be further enhanced by considering the parts of speech information in enhancing the correlation of the search items. Various topic patterns approach can be investigated further to provide better search results in a sequential manner. Discovering data patterns that may occur periodically for a search keyword. Further enhancing the search by mining the user behavior analysis in the website.

Received: April 2, 2017

References

[1] T. Cheng, X. Yan, and K. C.-C. Chang, “Supporting entity search: a large-scale prototype search engine”, In: *Proc. of the ACM SIGMOD International Conf. on Management of data*, Beijing, China, pp.1144-1146, 2007.

[2] L. Bing, W. Lam, T.L. Wong, and S. Jameel, “Web query reformulation via joint modeling of latent topic dependency and term context”, *ACM Transactions on Information Systems*, Vol.33, No.6, 2015.

[3] L. Li, L. Zhong, Z. Yang, and M. Kitsuregawa, “Qubic: An adaptive approach to query-based recommendation”, *Journal of Intelligent Information Systems*, Vol.40, No.3, pp. 555–587, 2013.

[4] V. Divya and C.R. Anju, “A survey on various summarization techniques”, In: *Proc. of the International Journal Of Engineering And Computer Science*, pp. 9528-9532, 2014.

[5] S.A. Babar and P.D. Patil, “Improving performance of text summarization”, In: *Proc. of the International Conference on Information and Communication Technologies*, Vol.46, pp. 354-363,2015.

[6] K. Balog, E. Meij, and M. de Rijke, “Entity search: building bridges between two worlds”, In: *Proc. of the 3rd International Semantic Search Workshop*, No.9, 2010.

[7] H. Zhang, M. Zhu, S. Shi, and J.R. Wen, “Employing topic models for pattern-based semantic class discovery,” In: *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL*, Vol.1, pp.459-467, 2009.

[8] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava, “Faceted search and browsing of audio content on spoken web”, In: *Proc. of the 19th ACM international conference on Information and knowledge management*, pp. 1029-1038, 2010.

[9] J. Pound, S. Paparizos, and P. Tsaparas, “Facet discovery for structured web search: A query-log mining approach”, In: *Proc. of the ACM SIGMOD International Conference on Management of data*, pp. 169-180, 2011.

[10] A. Casali, C. Deco, and S. Beltramone, “An assistant to populate repositories: gathering educational digital objects and metadata extraction”, *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, Vol. 11, pp. 87 – 94, 2016.

[11] L. Dong, K. Wu, and G. Tang, “A data centric approach to quality estimation of role mining results”, *IEEE Transactions on Information Forensics and Security*, Vol.11, No.12, pp. 2678 – 2692, 2016.

[12] A. Jain, R. Sharma, G. Dixit, and V. Tomar, “Page ranking algorithms in Web mining, limitations of existing methods and a new method for indexing Web pages”, In: *Proc. of the International Conference on Communication Systems and Network Technologies*, pp. 640-645, 2013.

[13] T. Kumari, A. Gupta, and A. Dixit, “Comparative study of page rank and weighted page rank algorithm”, *International Journal of Innovative Research in Computer and Communication Engineering*, Vol.2, No. 2, 2014.

[14] H.V. Reddy, P. Agrawal, S.V. Raju, “Data labeling method based on cluster purity using relative rough entropy for categorical data clustering”, In: *Proc. of the International Conference on Advances in Computing, Communications and Informatics*, 2013.

[15] C. Truica, F. Radulescu, and A. Boicea, “Comparing different term weighting schemas for topic modeling,” In: *Proc. of the 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing,* 2016.

[16] S. Shi and Q. Liu, “Improved normalized mutual information feature selection method in software cost estimation”, In: *Proc. of the 2013 International Conference on Information System and Engineering Management*, pp. 720-725, 2013.

[17] L. Li, J. Wu, and S. Zhu, “Implication intensity: Randomized F-measure for cluster evaluation” In: *Proc. of the 6th International Conference on Service Systems and Service Management*, 2009.

[18] W. Kong and J. Allan, “Extracting query facets from search results,” *In: Proc. of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pp. 93-102, 2013.