Mining Query Facets from the Search Results

A.Mahalakshmi, T.Yawanikha, D.Bhanu, V.P.Arul Kumar, K.M.Murugesan

Abstract— **Mining Query Facets includes multiple groups of words or phrases that are obtained from the query given in the search engine. The problem in addressing one important aspects of query facets related to the query given in the search engine is solved by QDMiner technique. The proposed system automatically extracts the query facets obtained from the top search results using the three HTML patterns. The main objective of the proposed system analyzes the problem in list duplication by aggregating the lists obtained in the search engine. The similarities between the extracted lists are estimated by k-means clustering technique which helps to avoid duplication among the websites. Mining facets will automatically rank all the lists extracted from the search results and display the higher priority lists in the top search results without any duplication and it also reduces the user the searching time to obtain better knowledge.**

Keywords—Query Facet, Higher Priority list, Duplication, Knowledge.

I. INTRODUCTION

Data mining is the powerful technology which analyze the data from different places and locations and that information is summarized into better knowledge. Data mining analyze the data stored in data warehouse. Data mining extracts the data from Data warehouse and the data are processed to evaluate patterns for providing useful knowledge to the user in an efficient manner. Data mining elicit the data using two learning approaches such as supervised and unsupervised. In supervised learning, relationship was recognized between dependent variable and explanatory variables. In unsupervised learning, there was no distinction between dependent variable and explanatory variable [1], [2].

Revised Manuscript Received on June 01, 2019.

A.Mahalakshmi, T.Yawanikha, D.Bhanu, V.PArulKumar, K.M.Murugesan, Assistant Professor, Professor, Assistant Professor, Professor, Department of Information Technology' ' Head - Department of Information Technology, Department of Computer Science and Engineering[5] Karpagam Institute of Technology - Coimbatore
mahalakshmi2626@gmail.com[1], yawanikha@gmail.com[2], bhanu.saran@gmail.com[3], thenmozhi2991@gmail.com[4], kmmcbe2010@gmail.com[5]

Searching query facets on search engine is an exploratory searching mechanism, which iteratively refine the search results based on faceted taxonomy. Faceted search analyze the models and framework to rank the facets based on exploratory searching mechanism. Faceted search is also known as faceted browsing which is a subset of faceted taxonomy [3].

The purpose of using web in day to day life is widely increased to access the information from web. A facet is a group of words or phrases which are summarized to provide one important aspects of the query facets related to the query given in the search engine. A query includes multiple facets that are processed and summarized into useful knowledge from different perspectives [4], [5], [8], [6]. Mining query facets are very important to extract and mine the knowledge obtained from the search results and they are ranked to obtain efficient search results without any duplication. Mining query facets from the search results saves the user's browsing effort and searching time in the search engine [21].

The rest of the paper focuses the following sections. Section II illustrates the related work. Section III explains the proposed work of mining query facets from the search results. Section IV describes the experimental result. Finally, Section V summarizes the conclusion of mining query facets and gives the direction for future work.

II. RELATED WORK

Mining query facets are implemented using various existing approaches. In this section, some of the related works which are used for the purpose of mining query facets from the search results.

A. Query Reformulation Technique

Query reformulation is the popular way to provide better search results. Query reformulation technique modifies the query based on user needs and expectations [7]. Search engine user logs are used to extract large amount of parallel data and snippet from the user clicked document. Query-to-snippet translation model is used to improve contextual query expansion. Translating queries into snippet only limited amount of phrases are extracted [8]. High precision rule based classifier organize the

Retrieval Number: A00440681S519/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

lists in order to find the similarities between the two lists to improve accuracy. High precision rule based classifier detects the reformulated query and it provide less beneficial to the user [7]. Query reformulation combines syntactic and semantic models to estimate the similarities among the lists and it provides less attention to the user [5].

B. Query Recommendation Technique

Query recommendation technique produce alternative query similar to the original query. Content based similarity model combines traditional approach to find the similarities between the lists to achieve high sparsity. K-means Neighbor method is used to cluster the lists to avoid duplication in the search results and it is hard to fix the k-value [10]. Query recommendation using clustering approach is used to detect similar queries to avoid duplication [11]. Queries are recommended in hidden web search engine to locate deep web sources. Query similarity calculator is used to estimate similarities among query keyword to avoid duplication [12].

C. Query Summarization Technique

Query summarization technique automatically summarizes the knowledge obtained from the search results by mapping the data into subset of simple descriptions [13]. Query summarization summarizes the description of the main topic of the text given in the search engine which provides relationship between the summary and the query.

Logic programming method summarizes the lists obtained from the search results. Tailored summarization technique is used to find the distance between the lists. Query summarization technique improves flexibility to obtain multi-document summary [13].

Query summarization using linguistic method examines and interprets the text to provide important information to the query which saves the users searching time [14]. Automated summarization technique compares with manual summarization technique to identify whether the sentences are ranked or not which helps to improve precision and it does not rank the higher priority lists [15].

D. Entity Search Technique

Entity search technique search the lists based on the entities, their attributes and the associated homepages. Entity search is a technique which provides relationship between the entities in the search engine and the lists are ranked to avoid duplication among the lists [16]. The queries are evaluated in both offline and online processing technique. In offline processing, the entities are extracted and the queries are globally processed.

In online processing, the queries are processed locally. Thus feasibility is improved in search results and it is hard to identify the entity and attribute value [17]. Information retrieval and semantic based web technologies are combined together to search the entities and their related attributes in the search engine thus performance is improved [18]. Ranking related entity rank the lists based on the following framework, 1. Co-occurrence Model: Co-occurrence technique improves performance and precision of search results. 2. Type Filtering: Type Filtering is used to categorize the queries. 3. Context Modeling: Context Modeling is a language model used to derive documents from source to target entity. 4. Homepage Finding: Homepage Finding is a technique which is used to rank the documents from homepage using standard language modeling approach. Thus precision is improved and error may occur while ranking [16].

E. Mining Facets and Faceted Search Technique

Mining Facets and Faceted search technique allows the user to examine the similar queries through multidimensional data [19]. Mining query facets extracts the set of lists obtained from the search engine based on the given query in obtaining better search results [20]. Automatic creation of Hierarchical faceted metadata is used to generate the metadata automatically from the textual descriptions of the items.

From the lexical database, large amount of information are manually modified to improve better search results and it provide less coverage of search results [20]. Unsupervised technique is used to extract the facets automatically from the search results. The important phrases along with context phrases are expanded in each document. The useful facets are extracted automatically from text databases based on their hierarchical level and their importance. Similarities between the lists are estimated by using frequency and rank based shifting algorithm [19].

III. THE PROPOSED METHOD

Mining query facets will automatically provide important aspects of the information given in the search engine. The problem in addressing one important query facets is solved automatically by QDMiner technique.

The main target of the proposed method recognize the problem in query facets obtained from the search results and provide one important query facets in website format without any duplication in the search results and further is reduces the user searching time but displaying higher ranked lists in the top search results.

Fig. 1 illustrates the proposed work of mining query facets from the search results. The proposed model includes 4

Retrieval Number: A00440681S519/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

modules. The first module describes list extraction from the search results.

The second module describes list weighting which rate the lists based on the positive and negative information. The third module describes list clustering to estimate the similarities among the lists and to avoid duplication and fourth module describes facet and item ranking to display the higher priority lists in the top search results.

In the proposed system, the query given in the search engine is given as input which is automatically reformulated and recommended to obtain one important query facets. The query facets obtained from the search results are automatically mined with the help of QDMiner technique from the HTML pattern.

By using three HTML patterns, such as free text pattern, HTML tag pattern and by using repeat region pattern, the lists are mined from the search engine.

Mining query facets will automatically extract and aggregate the lists obtained from the search results.

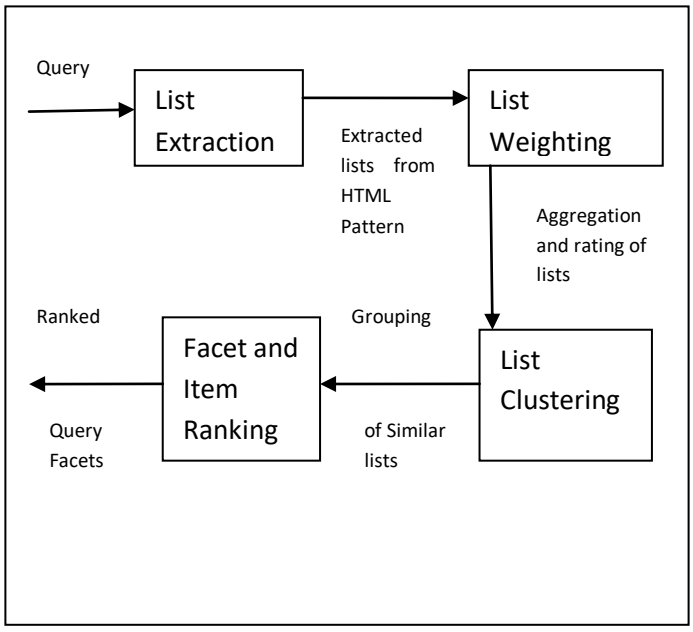

Fig.1 The Proposed Architecture

The lists obtained from the top search results are rated and ranked to avoid duplication among the lists and generate higher ranked lists in the top search engine in an efficient manner.

A. List Extraction

In list extraction technique, the query facets obtained from each document 'd' from the set of 'R' search result, the set of lists Ld={l'} are mined based on three HTML patterns.

1) **Free Text Pattern: extracts all texts within document d and the texts are splitted into sentences to employ the pattern items such as {, item}* (and | or) {other} from each sentence. These sentences are named as $TEXT_S$. A pattern {^item (: | -). +$} + extracts the lists from semi-structured paragraph which consists of two parts separated by a dash or a colon and these sentences are named as $TEXT_P$.**

2) **HTML Tag Pattern: extracts the lists using different HTML tags such as SELECT, UL, OL, TABLE and these patterns are named as $HTML_{TAG}$.**

3) **Repeat Region Pattern: To extract lists, the web pages that are repeatly recognized that are detected using vision-based DOM trees technique.**

B. List Weighting

The extracted lists from the search results may be informative or non-informative to the user. List weighting will rate the lists based on the positive and negative lists which are obtained from the list extraction. Usually, a positive list includes items that are more important to the query given in the search engine. Therefore, it is necessary to estimate the importance of each unique list 'l' by using the following components:

1) **S_{DOC}: Document Matching Weight (DMV): The items which contains positive lists must be periodically displayed in the top search results. So, the positive lists are rated by using (1),**

$$S_{DOC}=\sum d \in R(s_d^m . s_d^r) \quad (1)$$

where, $(s_d^m . s_d^r)$ represents the supporting score for each search result 'd'. s_d^m represents the percentage of items contained in document 'd'. s_d^r measures all the positive lists in the document 'd', $s_d^r = 1/\sqrt{rank_d}$, where $rank_d$ is the rank of each list contained in the document 'd'.

2) **S_{IDF}: Average Invert Document**

Retrieval Number: A00440681S519/19©BEIESP
Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

Frequency (IDF): Items which contains negative lists are non-informative to the query given in the search engine. So, the negative lists are rated by using (2),

$$S_{IDF}=\frac{1}{l}.\sum e \in l(idf_e) \quad (2)$$

where, $idf_e=\log((N-N_e+0.5)/(N_e+0.5))$, N_e represents the total number negative lists present in the document 'd' and N represents the total number of document obtained in the search results.

Finally, the lists that are rated based on the above techniques are aggregated to evaluate the importance of the list 'l' by using (3),

$$S_l=S_{DOC}.S_{IDF} \quad (3)$$

By rating the extracted lists in the above weighting methods, automatically the positive lists with high weight will be displayed in the top search results and the negative lists with low weight will be displayed at the last page of the search results.

C. List Clustering

In list clustering, the similar lists are grouped together to generate a facet related to a query. If two lists share enough information, then they are grouped together to estimate the distance $d_l(l_1,l_2)$ between the two lists l_1 and l_2 by using (4),

$$d_l(l_1, l_2) = 1 - \frac{|l_1 \cap l_2|}{\min\{|l_1|,|l_2|\}} \quad (4)$$

where, $|l_1 \cap l_2|$ represents the number of items that are shared within l_1 and l_2 . K-means clustering algorithm is used to group the lists based on the lists weight and thus it improves the quality of clustering and avoids duplication based on the following process,

(1). Let X={x1, x2… xn} be the set of data points and V= {v1, v2… vc} be the set of centers.

(2). Randomly select 'c' as cluster centers.

(3). Calculate the distance between each data points and cluster centers.

(4). Assign data point to the cluster center whose diameter from the cluster center is minimum of all the cluster centers.

(5). Recalculate the distance between each data point and new obtained cluster centers.

(6). If no data point was reassigned the stop, otherwise repeat from step 3.

D. Facet and Item Ranking

Facet and item ranking is used to display the higher priority lists in the top search results without any duplication and also it provides one important aspects of the query facet related to the query given in the search engine. Facet and item ranking will rank the lists obtained from the search results by using (5),

$$S_c = \sum_{G \in C(c)} S_G = \sum_{G \in C(c)} \max_{l \in G} S_l \quad (5)$$

where, S_c defines the importance of facet c, $C(c)$ represents the independent group of lists contained in the query facet c. S_G refers to the weight of the group of lists G and S_l refers to the weight of list l in the group G.

E. Performance Evaluation

Mining query facet technique will automatically extract the lists obtained from the search results and display the higher priority lists in the top search results without any duplication. The performance of query facets is measured based on nDCG, rp-nDCG, fp-nDCG, RI and FI.

In Fig.2, nDCG, rp-nDCG, fp-nDCG and RI are the existing parameters used to measure the query facets.nDCG (Normalized Discounted Cumulative Gain) is used to improve the quality of lists occurred in top search results.

In Fig.2, nDCG, rp-nDCG, fp-nDCG and RI are the existing parameters used to measure the query facets. nDCG (Normalized Discounted Cumulative Gain) improves the quality of lists occurred in top search results.

rp-nDCG(recall and purity aware-nDCG) calculates the entire output facet obtained from the search result. rp-nDCG parameter ranges from 0 to 1. If the value is closer to 1, then it denotes that the items are correctly classified into right query facets.

fp-nDCG(purity aware n-DCG based on first appearance of each class) measures the query facets by considering the purity of each facet that are multiplied by the correctly classified query facets.

RI (Random Index) measures all the facets without removing the duplication in an effective manner. FI (Faceted Index) measures the similarities among the query facets and avoids duplication with high precision.

In Fig.2, the blue color indicates the result of query facets obtained from HTML tag pattern and the green color represents the result of query facets obtained from the free text pattern.

In order to save the user's browsing time, the brown color

Retrieval Number: A00440681S519/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

will display all the search results. Faceted Index provides important query facets related to the query given in the search engine without duplication.

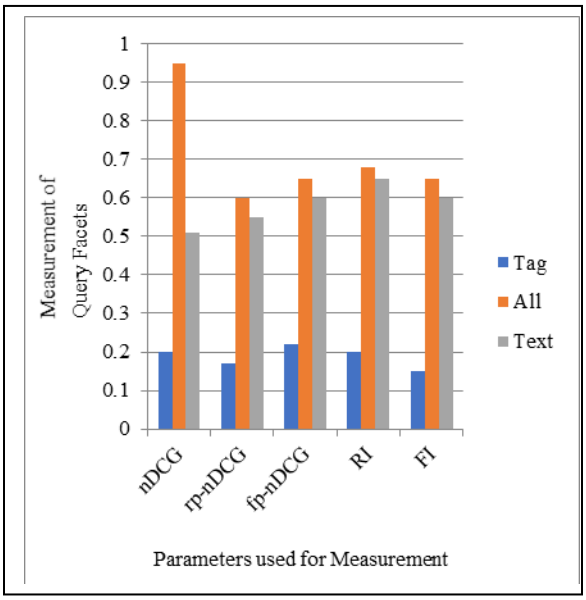

Fig.2 Performance Graph for measuring Query Facets

Table I shows the Statistics about the query facets generated with 25 search results, Dia_{max}=0.6 and W_{min}=0.3

TABLE I. PERFORMANCE EVALUATION

Description	**User Query**	**R and Query**
#queries	6.5	11.3
#results per query	8.9	12.6
#lists per document	3.6	2.3
#items per list	2.1	2.6
#Facets per query	18	22.6
#list per facet	3	2.5
#items/qualified items per facet	15.3/3.5	12.3/2.6
#good/fair facets among top 5 facets	1.5/1.1	0.8/0.1

In Table I, the column R and Query represents the existing result of generating query facets with 25 search results without removing the duplication of lists with high quality and they generate the query facets by multiplying the percentage of query facet. User Query represents the proposed work of generating one important query facets related to the query given in the search engine without duplication.

When the query given in the search engine, lists of query facets are obtained in the search results. These query facets are extracted to rate the lists based on their importance. After rating, in clustering similarities among the lists are evaluated and the duplication is removed. In table.1, Dia_{max}=0.6 is considered as maximum diameter and W_{min}=0.3 is considered as minimum weighting assigned to the list for estimating the distance based on K-means clustering technique between the two lists. While determining the similarities among the lists, if any two lists share same items then they are grouped as one and provide useful query without duplication, i.e. facets per query in user query as 18 represents in providing useful query facets without duplication. Then, the lists are ranked to display the higher priority lists in the top search result and while displaying the query facets if the value is closer 1 then it represents the good query facets otherwise it denoted as fair query facets.

IV. EXPERIMENTAL RESULT

To address the problem in finding the query facets and to provide one important query facets without duplication, first give query as an input in the search engine. After giving the query as an input in the search engine, the list of facets will be obtained in the search results that are related to the query. After obtaining the query facets from the search result, the lists are extracted in website format.

Retrieval Number: A00440681S519/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

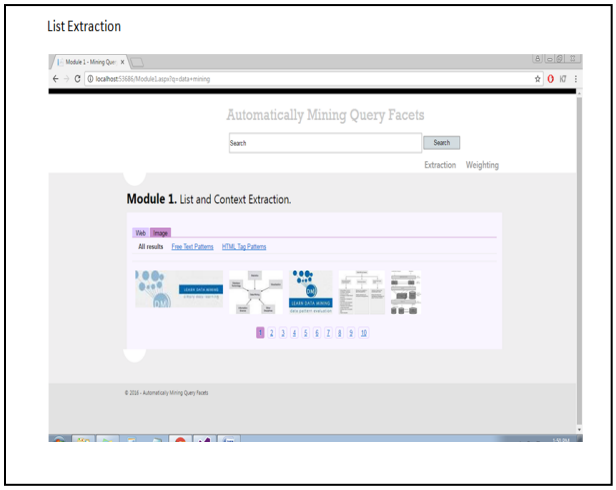

Fig.3 Image Extraction

The fig.3 shows the result of image extraction in the top

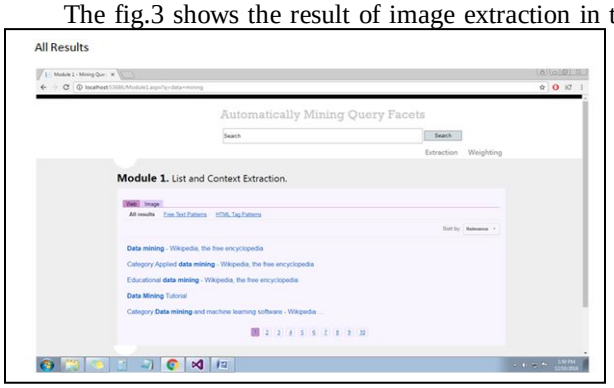

Fig.4 List Extraction

The fig.4 shows the result of list extraction in the top search results by using HTML patterns in the website format.

After extracting the lists obtained from the search results, the lists are rated based on their positive and negative list items.

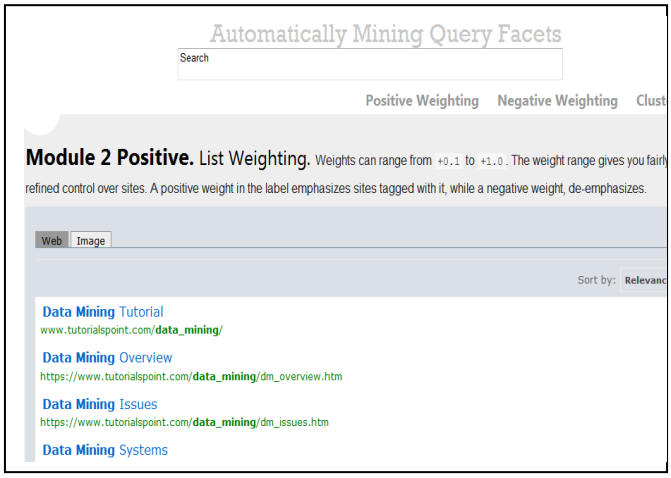

Fig.5 Positive Weighting Result

The fig.5 shows the positive weight of lists extracted from the search results based on Document Matching Weighting technique.

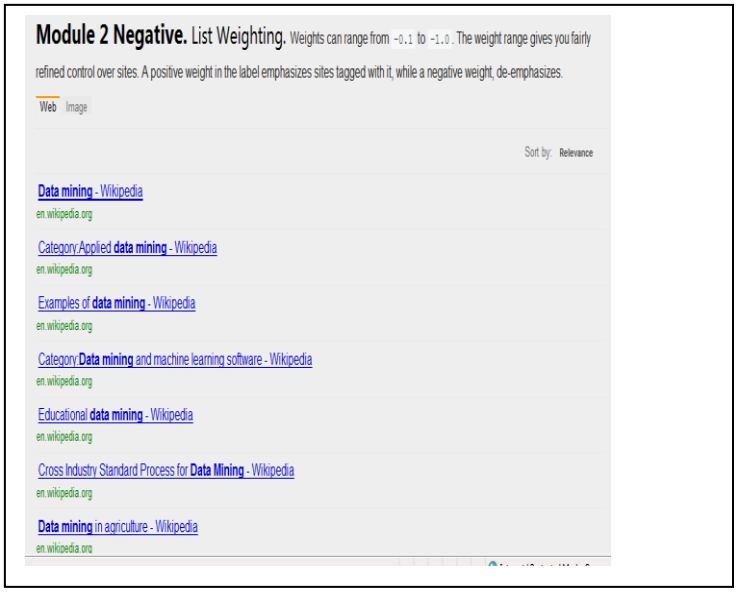

Fig.6 Negative Weighting Result

The fig.6 shows the negative weights of lists extracted from the search results based on Average Invert Document Frequency technique.

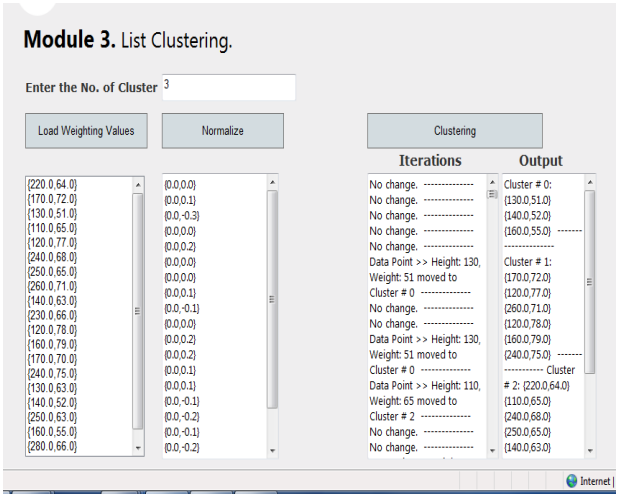

Fig.7 List Clustering

Retrieval Number: A00440681S519/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

In fig.7, the similarities among the lists are estimated using k-means clustering technique. Based on the given k values the lists are grouped and while grouping, if any two lists shares same items then it is grouped into one to avoid duplication.

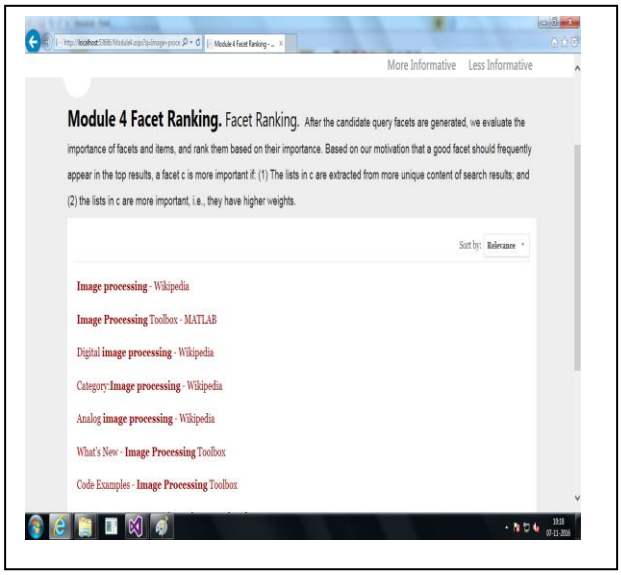

Fig.8 Higher Priority Facet Ranking

The fig.8 shows the ranking result of higher priority lists in the top search results without duplication.

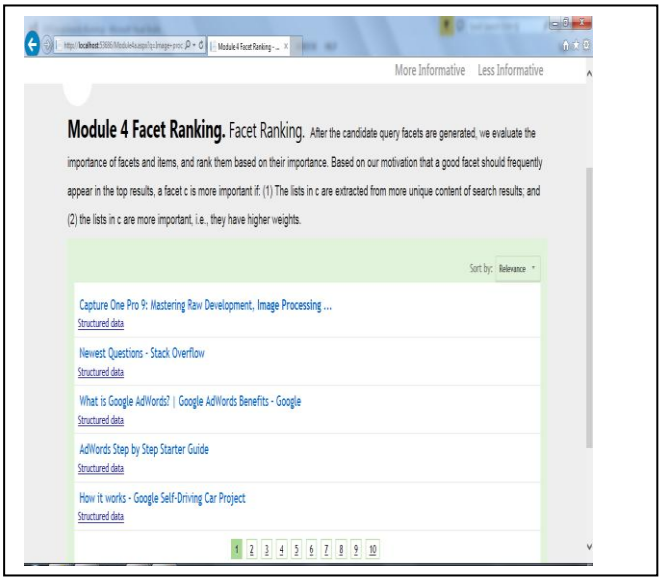

Fig.9 Lower Priority Facet Ranking

The fig.9 represents the ranking result of lower priority lists and thus the lower priority lists is displayed in the last page of the search results without duplication among the websites. Thus user's searching time will be reduced.

V. CONCLUSION AND FUTURE ENHANCEMENT

The problem in addressing one important aspects of query facets related to query given in the search engine are analyzed and solved by QDMiner technique. The proposed system provides a systematic solution in mining the query facets obtained from the top search results which aggregates the frequent lists from the search engine using free text, HTML tag and repeat region pattern of the HTML document automatically. After list extraction technique, the obtained lists are rated based on their positive weight and negative weight. The proposed system analyzed the problem in list duplication during the aggregation of lists in the search result. The quality of query facets is improved by estimating the similarities among the websites. The duplication among the lists is avoided by using k-means clustering algorithm, which automatically calculates the distance among multiple websites. While grouping the lists by using k-means clustering technique, if the minimum distance of the cluster node is not covered by a set of lists then the lists have lower priority and they are not displayed in the top search results. Finally by ranking the list, the higher priority lists will automatically displayed in the top search results without any duplication.

In future, instead of HTML pattern, a semi-supervised bootstrapping algorithm can be used for list extraction, which iteratively extracts more lists obtained from the top search results. Specific website wrappers can be used to provide high quality among the websites and by checking the homogeneity of lists, the quality of query facets can be further improved.

REFERENCES

[1] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar," A Brief Overview on Data Mining Survey", in International Journal of Computer Technology and Electronics Engineering. Volume 1, Issue 3.

[2] Anand V. Saurkar, Vaibhav Bhujade,Priti Bhagat, Amit Khaparde,"A Review Paper on Various Data Mining Techniques", in International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 4,April2014.

[3] Bifan Wei, Jun Liu, Qinghua Zheng, Wei Zhang,Xiaoyu Fu, Boqin Feng," A Survey Of Faceted Search", in Journal of Web Engineering,vol.12,no.1&2(2013)041-064.

[4] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava, "Faceted search and browsing of audio content on spoken web," in Proc. 19th ACM Int. Conf.
Inf. Knowl. Manage,

Retrieval Number: A00440681S519/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication

2010, pp. 1029–1038.

[5] A. Herdagdelen, M. Ciaramita, D. Mahler, M. Holmqvist, K. Hall,S. Riezler, and E. Alfonseca, "Generalized syntactic and semantic models of query reformulation," in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. retrieval, 2010, pp. 283–290.

[6] Stoica.E and M. A. Hearst, "Automating creation of hierarchical faceted metadata structures," in Proc. Human Lang. Technol. Conf., 2007, pp. 244–251.

[7] Huang.J and Efthimiadis.E.N, "Analyzing and evaluating query reformulation strategies in web search logs," in Proc. 18th ACM Conf. Inf. Knowl. Manage. 2009, pp. 77–86.

[8] Riezler.S, Liu.Y, and Vasserman.A, "Translating queries into snippets for improved query expansion," in Proc. 22nd Int. Conf. Comput. Ling., 2008, pp. 737–744.

[9] Zhang.Z and Nasraoui.O, "Mining search engine query logs for query recommendation," in Proc. 15th Int. Conf. World Wide Web, 2006, pp. 1039–1040.

[10] Hamada M.Zahera, Gamal F. El Hady, Waiel.F Abd El-Wahed, "Query Recommendation for Improving Search Engine Results", in Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I.WCECS 2010, October 20-22, 2010, San Francisco, USA.

[11] Khushboo Gulati, Narender," Query Recommendation in Hidden Web Search Engine using Web Log Mining Techniques", in International Journal of Computer Applications, Volume 102– No.13, September 2014.

[12] Damova.M and Koychev.I, "Query-based summarization: A survey," in Proc. S3T, 2010, pp. 142–146.

[13] Gupta, Gurpreet Singh Lehal Vishal," A Survey of Text Summarization Extractive Techniques", in Journal Of Emerging Technologies in Web Intelligence,vol.2,no.3,Aug 2010.

[14] Sherif Elfayoumy, Jenny Thoppil," A Survey of Unstructured Text Summarization Techniques", in International Journal of Advanced Computer Science and Applications, Vol. 5, No. 4, 2014.

[15] Bron.M, Balog.K, and Rijke.M, "Ranking related entities: Components and analyses," in Proc. ACM Int. Conf. Inf. Knowl. Manage, 2010, pp. 1079–1088.

[16] Cheng.T, Yan.X, and Chang.K.C, "Supporting entity search: A large-scale prototype search engine," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2007, pp. 1144–1146.

[17] Balog.K, Meij.E, and Rijke.M, "Entity search: Building bridges between two worlds," in Proc. 3rd Int. Semantic Search Workshop, 2010, pp. 9:1–9:5.

[18] Dakka.W and Ipeirotis.P.G, "Automatic extraction of useful facet hierarchies from text databases," in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 466–475.

[19] Stoica.E and Hearst.M.A, "Automating creation of hierarchical faceted metadata structures," in Proc. Human Lang. Technol. Conf., 2007, pp. 244–251.

[20] Megha R. Sisode, Ujwala M. Patil," A Survey on Query Recommendation Techniques and Evaluation of Snippet based Query Recommendation", in International Journal of Computer Applications (0975 – 8887) National Conference on Emerging Trends in Information Technology (NCETIT-2014).

Retrieval Number: A00440681S519/19©BEIESP

Published By:
Blue Eyes Intelligence Engineering &
Sciences Publication