ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

Mining Query Facets using Reverse Searching Algorithm

Pooja Mehesare, Prof.(Ms). D. V. Gore

Department of Computer Engineering, P.E.S. Modern College of Engineering, Pune, India

Abstract: Now a days everyone in the world searches using the search engine, but that will take more time for search result because it will not give the relevant results in one search. The user has to sort the number of links and after more searching the user gets the relevant link about his search. In this case user has to give more time for searching. So, for overcoming this problem facets is used. Facets is nothing but the relevant links about the query which user has given as an input. For finding the facets QDMiner method is used. QDMiner removes the duplicate contents and gives the relevant links about the search result. QDMiner has one disadvantage that it was not remove total irrelevant links so for improving QDMiner reverse searching algorithm is used. The reverse searching algorithm is proposed in the system for removing the irrelevant links from the seed collection and recommend link to the friend. Trust based analysis shows the graph that how many links are recommended.

KEYWORDS: Faceted Search, Query Facet, QD Miner Method, QT Clustering Algorithm, Reverse Searching Algorithm, Recommend Link, Trust Based Analysis

I. INTRODUCTION

The query facet is a set of items which provide knowledge aboutquery,inthatitemcontainwordorphrases sousercansee the query facets with original results with different way, thus user can understand the detail important information about query without browsing the multiple pages. The user gives a query then they get the multiple links in that many links contains the same contents. By using one of the word in the user query it will provide multiple relevant and irrelevant links to the user because of that the user take more time to search the relevant link for their search result and also they are not get relevant link. So for overcoming this problem we use the query facets, with the help of query facets user can save the browsing time and also get the relevant link about search query. The QDMiner method is used to find the relevant links of the search result. In that query is take as an input then multiple steps of QDMiner is applied on the query such as list extraction, list weighting, list clustering and ranking. For the list clustering QT Clustering algorithm is used. QDMiner is not remove all irrelevant links so for improving the QDMiner reverse searching algorithm is used. The reverse searching algorithm is used to remove the irrelevant links from the seed collection so that the user easily get the facets. When the facets is generated one of the link in that facets is recommended totheir friend. Trust based analysis shows the graph that how many links are recommended. In reverse searching algorithm URL is relevant to query or not is find out by using two ways first is if the page contains relevant result to the query then it is relevant. Second is if the fetch web sites pages is larger than the user query then also that URL page is relevant to the query.

II. REVIEW OF LITERATURE

Many methods are used to generate the facets, but it will take more time to generate the facets. So, for overcoming this problem QDMiner is used.
QDMiner gives the query facets from the search results. In generalcasessomelinkshavedifferentURLbutinthatcontent are same that is duplicate content so QDMiner finds main content from different documents and remove the duplicate content get the relevant links that is the facets. There are four steps of QDMiner that is list extraction, list weighting, list clustering and item ranking. For the list extraction three patterns are used such as free text, HTML tags, and repeat

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

regions, for the list weighting two components are used document matching weight ,average invert document frequency ,and for the list clustering QT clustering algorithm is used ,after getting the similar list ranked that lists and then get the facets [1].

Reverse searching algorithm is used to remove the irrelevant links. The URL is relevant to query or not is find out by using two ways first is if the page contains relevant result to the query then it is relevant. Second is if the fetch web sites pages is larger than the user query then also that URL page is relevant to the query [2].

Generally the facets is used in the e-commerce and other application,theideaoffacetsisnotexploreforthegeneralweb search. For taking the user feedback into document ranking they investigate the Boolean filtering and soft ranking models. This is help the user to find the relevant search and given the information about their subtopics [3].

For recognizing query facets they developed a supervised approach based on the graphical model. Two algorithms are proposed in this paper. Their evaluation is combines the precision and recall with the grouping quality of the facets [4].

In this paper four components are analysis that is the co-occurrences models, type filtering, context modelling and homepage finding. Their initial focus on the recall. In this paper recall has a very high score and also providing a solid starting point for expanding our focus to improve precision [5].

In this paper they shows the improving recommendation for the long queries by using the templates, many times user give long queries as an input in that system have to search the query by using template that means taking any one important word in the query and give the relevant link to the user [6].

In this paper the facets are recommended to the user but in this the web search is structured. For presenting a structured data the facets is used as a tool for navigating, refining and grouping the result [7].

Now a days spoken web is introduced it is a web of Voice Sites and it is accessed in phone. In this paper they apply the facet search and web spoken problem. It is also provide the rank mechanism to rank the facets from their search result. In this paper they introduced method is the first method of facets for the audio search [8].

In this paper the traditional faceted search is extended. In which first extension is adds flexible, dynamic business intelligence aggregation to the faceted application and the second extension is shows that how one can efficiently extend a faceted search engine to support correlated facets [9].

In this paper facets are automatically extracted from the text databases. For the automatic extraction of facets unsupervised method is used. For constructing the browsing the facets the terms are compare in the original databases and the expanding databases [10].

III. SYSTEM ARCHITECTURE / SYSTEM OVERVIEW

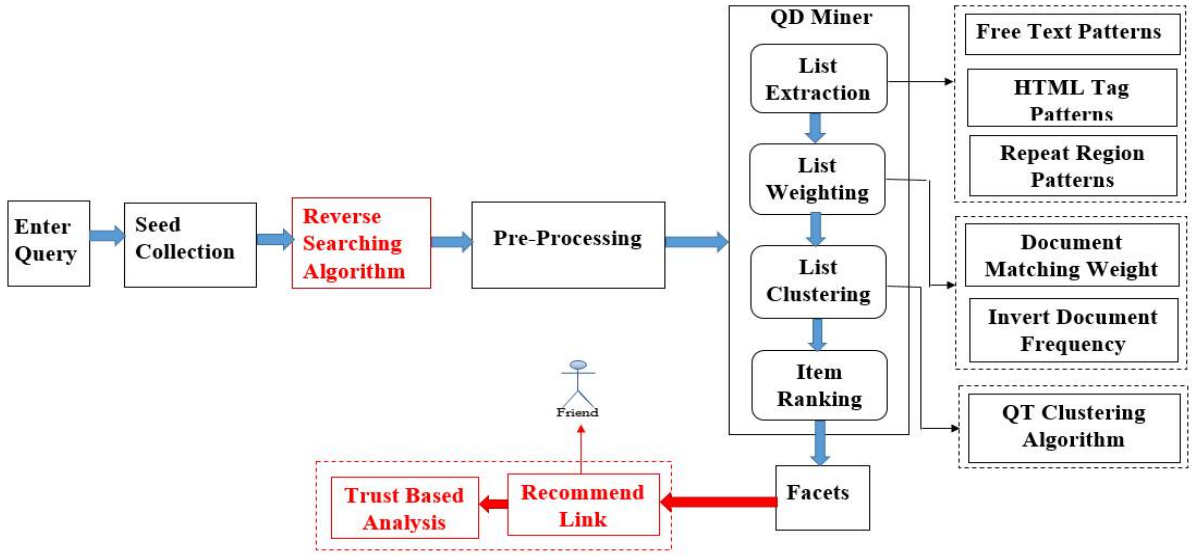

Fig. 1. Block Diagram of Proposed System

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

1) System Overview: The system is for the QDMiner method with reverse searching algorithm. The QDMiner is used to generate the query facets for the search results and reverse searching algorithm is used to remove irrelevant links from seed collection. In this system the query is given as an input to the system then the number of links are generated for that query that is nothing but the seed collection. In that some links are relevant and some are irrelevant. The reverse searching algorithm is improved this system which is remove the irrelevant links in the seed collection. Then pre-processing is there and QDMiner method is applied on the system. In QDMiner there are four steps such as list extraction, list weighting, list clustering and item ranking. In list extraction QDminer extract the list and their context from each search result documents based on three patterns such as free text patterns, HTML tag patterns and repeat region patterns. After list extraction the weight of list is calculated in that two components are used that is document matching weight and average invert document frequency. After list extractionlist clustering isperforminthatthereisthegrouping of similar list and for the clustering QT Clustering algorithm is used. After getting the similar list we ranked that list by using the item ranking. After all this steps system generate the facets that one of the link in that facets is recommend to friend and also check recommended links in graph using trust based analysis.

2) Mathematical Model:
Let S be the System as,

$$S = \{I,F,O\}$$

where,
I= Query q and retrieved top search results of query from search engine.
O=Facets generated for a query q.
F= {F1, F2}
where
F1 is the function for QDMiner Method
F2 is the function for Reverse Searching Algorithm

F1 is QDminer Method

- List Extraction:
  - Input=Query.
  - Output=Set of list is extracted.
  - We use three different types of pattern to extract list:
    - Free text pattern (Text).....(1)
    - HTML tag patterns (Tag).....(2)
    - Repeat region patterns (RepeatRegion).....(3)
- List Weighting:
  - Input=Extracted list.
  - Output=List weighting.
  - Final Weighting List:

$$S_l = S_{Doc} \cdot S_{IDF} \quad \text{.....(4)}$$

$$S_{DOC} = \sum_{d \varepsilon R} (S_d^m \times S_d^r) \quad \text{.....(5)}$$

$$S_{IDF} = \frac{1}{|l|} \times \sum_{e \varepsilon l} idf_c \quad \text{.....(6)}$$

$$idf_c = log \frac{N - N_e + 0.5}{N_e + 0.5} \quad \text{.....(7)}$$

where,
S_l is the weighting list.
S_{Doc} is the document matching weight.
S_{IDF} is the average invert document frequency.
$S^m_d \times S^r_d$ is the supporting score by each result d.
S^m_d is the percentage of items contained in d.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

S^r_d is measure the importance of document d.
N_e is the total number of documents that contain item e in the corpus.

- List Clustering:
  - Input=List Weighting.
  - Output=Groups data into high quality cluster.

$$w_c = |Sites(c)| \quad \text{.....(8)}$$

where,
Sites(c) is the set of websites that contain the list in c.
$w_c$is the weight of cluster.

- Item Ranking:
  - Input=Groups of Data.
  - Output=Facets and Item Rank.

$$AvgRank_{c,e,G} = \frac{1}{|L(c,e,G)|} \sum_{l \epsilon L(c,e,G)} rank_{e|l} \quad \text{.....(9)}$$

where,
$AvgRank_{c,e,G}$is the average rank of item e within all list extracted from group G.
L(c,e,G) is the set of all lists in c.
Ranke|l is the rank of an item e within a list l.

F2 is the function for Reverse Searching Algorithm

- Reverse Searching Algorithm:
  - Input=Seed Sites.
  - Output=Relevant Sites.

$$\bar{FSS} = \{U, A, T\} \quad \text{.....(10)}$$

where,
U,A,T is the vectors corresponding to the feature context of URL, anchor, and text around URL.

IV. ALGORITHMIC APPROACH

QT Clustering Steps:

- Initialize the threshold distance allowed for clusters and the minimum cluster size.
- Build a candidate cluster for each data point by including the closest point, the next closest, and so on, until the distance of the cluster surpasses the threshold.
- Save the candidate cluster with the most points as the first true cluster, and remove all points in the cluster from further consideration.
- Repeat with the reduced set of points until no more cluster can be formed having the minimum cluster size.

Reverse Searching Steps:

- Input: Seed sites.
- Output: Relevant sites.
- while candidate site do.
- // Pick a deep website.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

- site = seedSiteCollection(siteDatabase, seedSites)
- links = extract links (link).
- page = compareUrl(link).
- relevant = classify (page).
- if relevant then.
- list.add(page).
- return list.
- end.

V. SYSTEM ANALYSIS

1) Dataset Information: The database which are used in this system is the online database of google by google API. API will be the GOOGLESEARCHURL = https://www.google.com/. Also use the Jsoup for the Java HTML Parser in that there are three methods such as DOM is used to navigate the documents and extract the data, CSS and jquery is the selector syntax to find out the matching elements.

2) Software and Hardware Requirements: The above system is implemented using 256 MB (min) RAM and above with Speed of 1.1 GHz. The processor used is Intel core i3 andabove.20GBHarddiskisusedforthispurpose.Operating System Windows 10 Pro with Java 1.7 is used. Tomcat 5.0/6.x server is used. Eclipse consider as the development environment. MySQL is used as database for implementation.

3) Performance Parameter:

A. fp-nDCG Purity aware nDCG based on the first appearance of each class.

$$w_i = \frac{\left|c_i^l \bigcap c_i\right|}{|c_i|} \quad \text{.....(11)}$$

B. rp-nDCG Recall and purity aware nDCG. rp-nDCG is calculated based on all output facets.

$$w_i = \frac{\left|c_i^l \bigcap c_i\right|}{|c_i|} \times \frac{\left|c_i^l \bigcap c_i\right|}{\left|c_i^l\right|} \quad \text{.....(12)}$$

where,

$\frac{\left|c_i^l \bigcap c_i\right|}{\left|c_i^l\right|}$ is the percentage of items in c_i^l matched by the current output facets c_i .

w_i is the weight for each automatic facet c_i.

C. Implemented Results

The graph shows the result of QDMiner. Graph has x-axis as a accuracy range and y-axis as a fp-nDCG and rp-nDCG.

1) Results: We give the query Shoes, Jeans and T-Shirt as a input to the system, we get the relevant links for the search result. From the above graphical and tabular results, we get the facets.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

TABLE I
RESULTS FOR QDMINER AND REVERSE SEARCHING ALGORITHM

Categories	QDMiner		Reverse Searching Algorithm	
	fp-nDCG	rp-nDCG	**fp-nDCG**	**rp-nDCG**
Shoes	10.8	0.91	**17.4**	**0.94**
Jeans	4.80	0.82	**8.00**	**0.88**
T-Shirt	3.87	0.79	**5.25**	**0.84**
Gold Bangles	2.71	0.73	**3.33**	**0.76**
Gold Rings	2.00	0.66	**5.00**	**0.88**

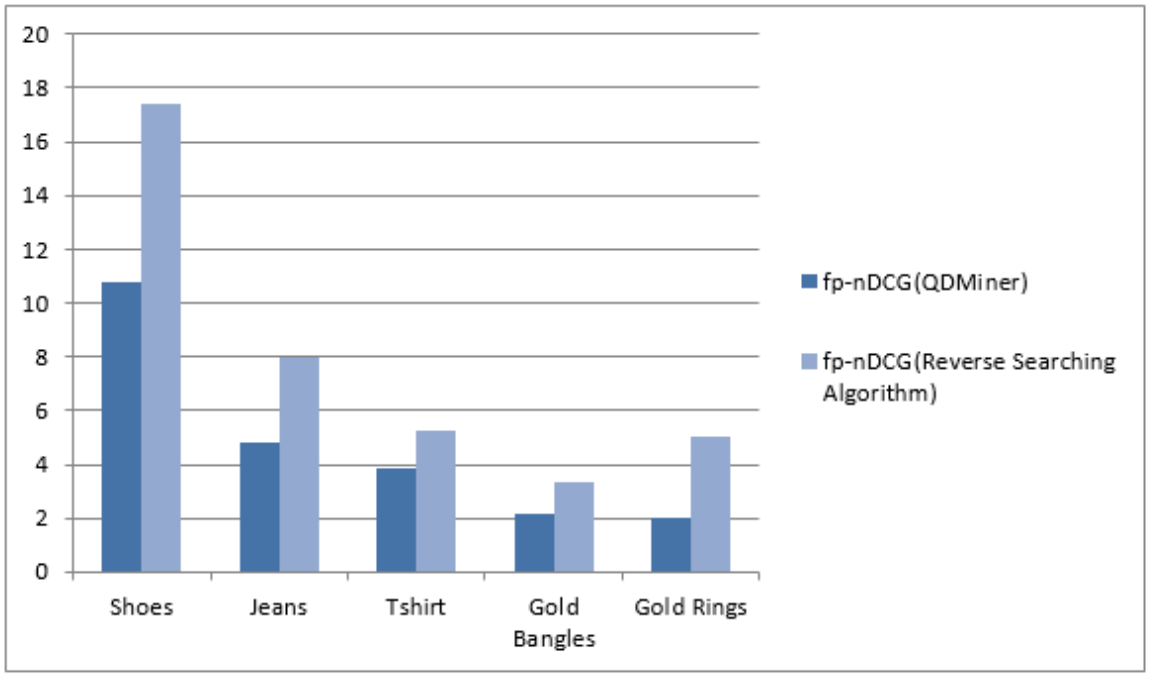

Fig. 2. Results Graph for Precision

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 7, July 2018

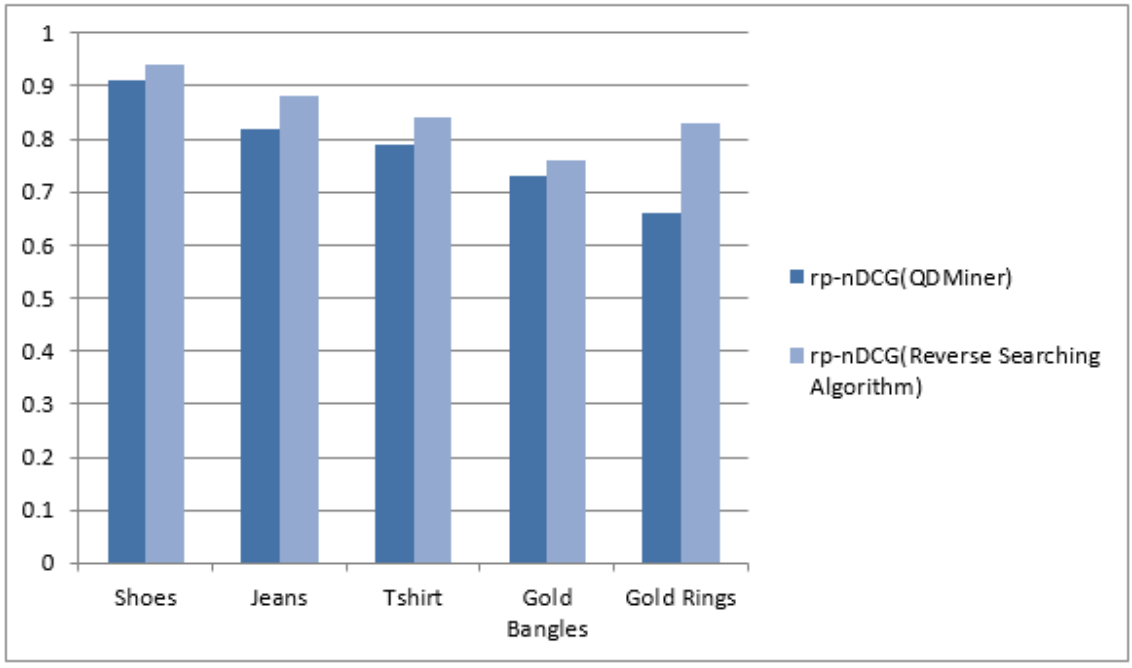

Fig. 3. Result Graph for Recall

VI.CONCLUSION

QDminer is used to generate the facets for a search results.Reverse searching algorithm is used to remove irrelevant links. Using reverse searching algorithm irrelevant links have been removed.For single word query precision is increased by 33.32%. For single word query recall is increased by 33.07%. For the double word query precision is increased by 50% .For the double word query recall is increased by 49.83%.The performance of the system has improved using reverse searching algorithm along with QDMiner.

REFERENCES

[1] Zhicheng Dou,Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song, "Automatically Mining Facets for Queries from Their Search Results,"in IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,Vol. 28, No. 2, February 2016.
[2] Feng Zhao, Jingyu Zhou, Change Nie, Heqing Huang and Hai Jin,"Smart Crawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web Interfaces,"Proc. ACM Int. Conf. Inf. Knowl. Manag., 2015.
[3] W. Kong and J. Allan,"Extending faceted search to the general web," Proc. ACM Int. Conf. Inf. Knowl. Manage., 2014.
[4] W. Kong and J. Allan,"Extracting query facets from search results,"Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013.
[5] I. Szpektor, A. Gionis, and Y. Maarek, "Improving recommendation for long-tail queries via templates," Proc. 20th Int. Conf. World Wide Web,pp. 4756., 2011.
[6] J. Pound, S. Paparizos, and P. Tsaparas ,"Facet discovery for structured web search: A query-log mining approach," Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011.
[7] M. Bron, K. Balog, and M. de Rijke,"Ranking related entities: Components and analyses,"Proc. ACM Int. Conf. Inf. Knowl. Manage, 2010.
[8] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava,"Faceted search and browsing of audio content on spoken web,"Proc. 19th ACM Int. Conf. Inf. Knowl. Manage., 2010.
[9] O. Ben-Yitzhak, N. Golbandi, N. HarEl, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev,"Beyond basic faceted search,"Proc. Int. Conf. Web Search Data Mining, 2008.
[10] blackW. Dakka and P. G. Ipeirotis, "Automatic extraction of useful facet hierarchies from text databases,"Proc. IEEE 24th Int. Conf. Data Eng., 2008.