Rich Search Experience for Query with Mining Facets

Arti P. Dhoot[1]* Dr. S. R. Todmal[2]

[1]Department Of Computer Engineering, JSPM's Imperial College of Engineering, and Research, Pune, India

[2]Department Of Computer Engineering, JSPM's Imperial College of Engineering and Research, Pune, India

Abstract – Query with Facet Mining mechanism helps user to search, locate and access appropriate data from online web contents. The system generates facets not only for given query but also for whole collection. In the paper, proposed system discovers to mechanically find query related feature of search for open-domain queries using search engine. Facets for a query are automatically pull out from top search list for web. It enables user to recognize query aspects which promote improved search experience of user. In this paper, proposed system retains facets output of the query in database that can be delivered to the user searching for same query .and this improves the performance of the system and user search experience. When a user relates with the system with query, the system creates Facets for his query and give back to the user. Also it stored the generated facets in a database table. When any random user request same query, the system will first check facets for query in database, and it returned if it is available otherwise it generates facets for new query. So it saves system processing time for duplicate query and improves the performance. System will display pre-query for user entered query according to past searched history.

Keywords- Clustering, faceted search, Query facet, QDMining, Ranking, Pre-query, Page parsing, summarization

----------X----------

1. INTRODUCTION

An effective approach for faceted search is the scope of its implementation. Query with Facet Mining mechanism helps user to search, locate and access appropriate data from online web contents. It is popularly used in e-commerce based application. An effective approach of facets search is implemented in the system. Facet search based system are mostly built on particular domain (for example: Product Search) or existing category of facets. Proposed facets searching system gathers information and media content in online search results. The system generates facets not only for given query but also for whole collection. In this paper, proposed system discovers to automatically find query related aspect of search queries using search engine. Facets from query are retrieved from the top web search results of the query automatically with no any additional domain knowledge. Query Facets are promising data sources that enable a research on open domain facet. When a user relates with the system with query, the system creates Facets for his query and give back to the user. Also it stored the generated facets in a database table. When any random user request same query, the system will first check facets for query in database, and it returned if it is available otherwise it generates facets for new query.

a. **Motivation:**

1. The challenges come from the large and diverse nature of the web, and it makes difficult to create and recommend facet.

2. The query facet contains a group of words and phrases that provides the information about query.

3. Previous models typically generate words and phrases related to the original query, but do not consider how these words and phrases would fit together in actual.

b. Objective and Scope:-

1. Automatically facet mining:-

Generating facets related to user expected query from their search result.

c. Goal:-

Recommend the facet to user according to user entered query.

2. REVIEW OF LITERATURE

Improving the Efficiency of Web Crawler by Integrating Pre Query Approach- Author L. Bing, W. Lam, T.-L. Wong, and S. Jameel present The amount of data consumed by crawler while searching is massive. The crawler hunts large amount of data which may contain lots of unrelated information. Also a lot of time is misused for examining related data among the huge amount of unrelated results got by crawler and user has to waste a time while crawling on web while scanning irrelevant links also. Pre/Post query processing approach and site-based searching approach can be combine in order to pre-processing the user query. By mixing of different processing approaches and link ranking approaches save a lot of valuable user time. Post query system may also filter out all irrelevant information which is not necessary according to the query which is been fired, and gives the expected results [11].

Automatically Mining Facets for Queries from Their Search Results-Author Zhengbao Jiang, Sha Hu, Ji Rong Wen, and Ruihua Song: Here, the focus was on pulling out facet from document based on query of the user. In this paper mining of facet is done through the list in HTML element which mainly contains important feature of query. There is automatic query facet mining through clustering of data. They perform experiments to evaluate the generated facets. They had built two data sets and applied metrics which were already exist and two more new combined metrics and evaluated the feature of query facet generated [10].

Search Result Diversification Based on Query Facets- Author Sha Hu, Zhi-Cheng Dou, Xiao-Jie Wang presents an approach of query subtopic, it covers embedding phrases, classification distribution of query. It also signifies different Semantic term in vector space model finds a similarity of query results clustering. It helps automatically looking for subtopics for a user query and which are returned in the form of string. It involves different resources like suggestion of query, top-ranked search results [3].

Combination of multiple semantics to perform Mining of Query Sub-Topic- Author Lizhen Liu, Wenbin Xu, Wei Song, Hanshi Wang and Chao Du represents facets of query based on user interest. Diversified search results for query where facets provides collection of words which further provides intent of query. The semantic outcome of query different combination of phrases helps to increase search experience of user [4].

Beyond basic faceted search Author O. Ben-Yitzhak, N. Golbandi, N. HarEl, R. Lempel present research was based on OLAP model to analyze data online as per user interest for mining query. OLAP contain relational database, so facet mining features was supported for free text queries using metadata [5].

Dynamic faceted search for discovery-driven analysis- Author D. Dash, J. Rao, N. Megiddo presents was based on attribute driven aspects of query. For a user query a dynamically set of attributes are chosen and combine with query to present to the user. It was similar to OLAP based analysis of aggregating values. It represents both textual content related to query and also structured attributes [6].

Extracting Query Facets from Search Results- Author Weize Kong and James Allan represents a graphical model technique to find query facets from search result. The graphical model helps to generate clusters of related terms in a query facet. It focuses on aggregating the two related worlds to obtain the dependencies. They used two algorithms for approximate conclusion on the graphical model. We planned a new evaluation metric for their task to combine recall and precision of facet generated [2].

Improving automatic query expansion Author M. Mitra, A. Singhal present the data extracted from a webpage hidden content through search engine. Here the data is taken from querying the interface not by hyperlinks and through reading from dynamically generated result page [1].

Locating hidden web entry points using adaptive Crawler Luciano Barbosa, and Juliana Freire, This research helps to search relevant data from the page content. It focuses on prioritizing links which are promising. It focuses on a technique for automatic searching of patterns of useful links. So it requires less manual setup. This paper also focus on extending the crawling process of determining and retrieving forms for specific database domain which is useful for learning online [8].

Harvesting Deep-web Interfaces using two stage crawler -Author Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin proposed a smart crawler which is a two staged crawler, retrieves data from web pages deeply. In the first step, smart crawler searches through website for hidden web pages using search engines, instead of visiting too many pages. Smart crawler also ranks web pages to retrieve highly related data which increases efficiency of crawled results. In the second step, Smart crawler uses adaptive link prioritizing to perform fast crawling to retrieve relevant links [7].

Searching Documents Based on Relevance and Type- Author presents Jun Xu1, Yunbo Cao1 presents a survey is based on types of models, that

www.ignited.in

are, relevance model and type model. The relevance model checks if the document is important for search query. The type model checks if the documents belongs to the retrieved or collected document type. Different experiments performed to check the effectiveness of typed search. They have performed testing for different domains and proved their results are consistent. This survey paper used BM25 and Logistic Regression techniques for both types of models[9].

3. PROPOSED SYSTEM APPROACH

Fig.1 presents While, entering query user will get pre-query result for entered query according to his past searched history. If query is first time then following process will execute. If queris second time log database will check and facets are display to user.

1) **Seed collection**: Here input to system is collect from online API. Which accepts the query and according to query it gives links according to query. After that reverse searching is performed to find seeds are relevant to query or not.

2) **Unique website identification**: Here unique URL Only finds and that unique only passes to next step. We were performing these step after getting seeds from seed collection by matching two pages' content. So for the next step of page parsing will not apply on duplicated links. That will save the time of our system. In this model, it is assumed that for one website, data can be duplicated in the list. But for different sites, the data can be different which add vote for facets weight. Again, sometimes, data list from two websites can also be duplicated. Also in case of mirror websites where domains are different but contents are same. Sometimes, content of one website can be republished in different websites so information might repeat in different sites. Also, some software may be used to publish information in different sites, that may result in duplicate content.

3) Page parsing process: In page parsing, list content is retrieved from different HTML elements for example HTML Table, UL, OL, SELECT following pattern .It contains facet and links that will display to user.

4) Query aspects from page: After performing page extraction we get facets and links. The SELECT HTML element contains 'OPTION' child tags form where text is used to form list. UL/OL also provides text content to form list.

5) Facet classification and ranking: Facets are clustered according to different classes. It cluster data of similar facets and rank the facets. Top search results should frequently display good facet.

6) Log database: After getting facets are stored in log file that will save users searching time from second time entered same query.

4. SYSTEM ARCHITECTURE

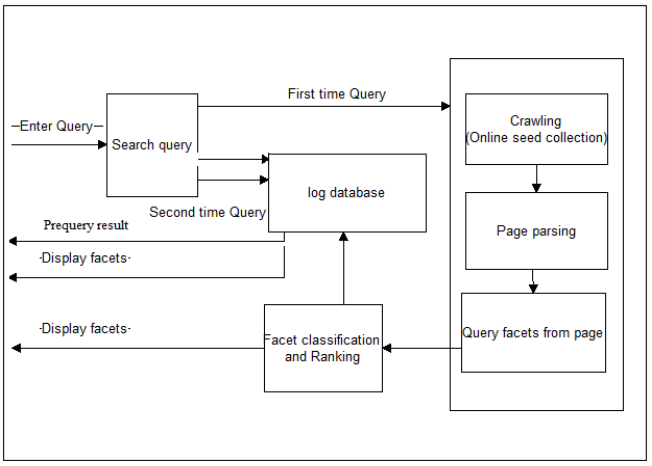


Fig 1: System Architecture

6. MATHEMATICAL MODEL

Notation:

It is contain importance of

S_d^m ----------percentage of items contained in d

s_d^r ----------measures the importance of document d.

Nl;d : the number of items which appear both in list l and document d,

jlj: the number of items contained in list l,

Equation:

$S_d^m = \frac{N_{L,d}}{|l|}$ ----------[1]

$s_d^r =$ 1/sqroot of rank doc ----------[2]

$S_{Doc} = \sum_{d \in R}(S_d^m \cdot s_d^r)$ ----------[3]

7. ALGORITHM

Algorithm for QD Miner

QD Miner:

Steps:

Input: Query Output: Facets related to query

Step1: Get Query from user

Step2: Check query facets are in log database. If present goto step 10 else goto step 3.

Step3: Seed collection by api

Step4:get relevant link

Step5: Extract facets from each link

Step6: Weight the facets

Step7: Cluster the facets

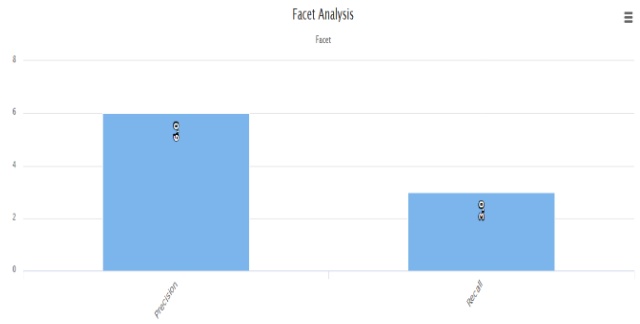
Step8: Rank the facets

Step9: Store facets in log database

Step10: Show facets to user.

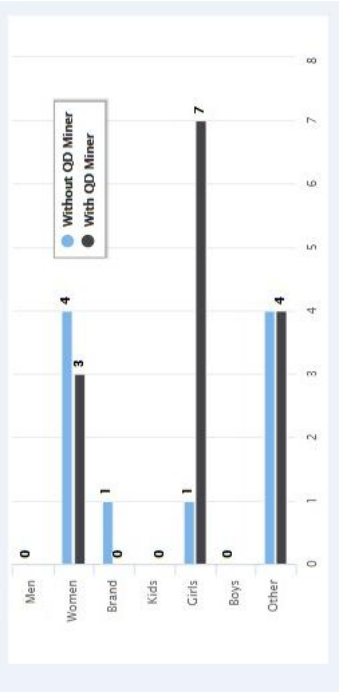
8. EXPERIMENTAL SET UP

We propose QD Miner to extract facet of user entered query.



Graph 01. Precision Recall for facets.

Explanation:



Graph 02: Comparision of without QD Miner with QD Miner

Explanation: Here facets are displayed according to only seed get from online database. With QD Miner shows no. of facets for each class by performing QD Miner.

9. CONCLUSION

In this paper, we study the problem of finding query facets comparatively faster through suggestion. We propose a systematic solution, which we refer to as QDMiner, to automatically mine query facets by aggregating frequent lists from free text, HTML tags, and repeat regions within top search results. We further analyze the problem of duplicated lists, and find that facets can be improved by modeling fine-grained similarities between lists within a facet by comparing their similarities. To improve performance, we are using log file of generated facets that are stored. This contribution improve time of searching over existing QD Miner. Also when gets the seed that seed are checked weather they are duplicate or not so its save the time to extract facet. Log database is maintained to reduce the users searching time for entered query facets. User gets Pre-query recommendation of query.

REFERENCES

Automatically Mining Facets for Queries from Their Search Results Zhicheng Dou,Member, IEEE, Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song

D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman (2008). Dynamic faceted search for

www.ignited.in

discovery-driven analysis, in ACM Int. Conf. Inf. Knowl. Manage., pp. 312.

Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang, Hai Jin, Smart Crawler (2015). A Two stage Crawler for Efficiently Harvesting Deep-Web Interfaces, in IEEE Transactions on Services Computing Volume: PP Year: 2015.

Jun Xu, Yunbo Cao, Hang Li, Nick Craswell and Yalou Huang (2007). Searching Documents Based on Relevance and Type, in ECIR 2007, LNCS 4425, pp. 629 636, 2007.

L. Bing, W. Lam, T.-L. Wong, and S. Jameel: Web query reformulation via joint modeling of latent topic dependency and term context.

Lizhen Liu, Wenbin Xu, Wei Song, Hanshi Wang and Chao Du: Query Subtopic Mining by Combining Multiple Semantics

Luciano Barbosa, and Juliana Freire (2007). An Adaptive Crawler for Locating HiddenWeb Entry Points, in May 812, 2007, Banff, Alberta, Canada. ACM 9781595936547/07/0005.

M. Mitra, A. Singhal, and C. Buckley (1998). Improving automatic query expansion, in Proc. 21st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1998, pp. 206-214.

O. Ben-Yitzhak, N. Golbandi, N. HarEl, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev (2008). Beyond basic faceted search, in Proc. Int. Conf. Web Search Data Mining, pp. 33-44.

Sha Hu, Zhi-Cheng Dou, Xiao-Jie Wang (2013). Search Result Diversification Based on Query Facets

Vishakha Shukla (2016). Improving the Efficiency of Web Crawler by Integrating Pre Query Approach, Year- 2016

Weize Kong and James Allan (2013). Center for Intelligent Information Retrieval, Extracting Query Facets from Search Results, in July 28August 1, 2013, Dublin, Ireland.

Corresponding Author

Arti P. Dhoot*

Department Of Computer Engineering, JSPM's Imperial College of Engineering, and Research, Pune, India

E-Mail – dhoot.arti@gmail.com