ISSN (Online): 2456-5717

International Journal of Advanced Research in Basic Engineering Sciences and Technology (IJARBEST)
Vol.3 Issue.8 August 2017

A Survey on Smart Web Query Searching Using Automatically Extracted Facets

Duhita Pawar[1], Prof. Vina M. Lomte[2]
[1, 2] RMD Sinhgad School of Engineering, Pune

Abstract— **In this paper, a detailed survey on different facet mining techniques, their advantages and disadvantages is carried out. Facets are any word or phrase which summarize an important aspect about the web query. Researchers proposed different efficient techniques which improves the user's web query search experiences magnificently. Users are happy when they find the relevant information to their query in the top results. The objectives of their research are:**
1) To present automated solution to derive the query facets by analyzing the text query.
2) To create taxonomy of query refinement strategies for efficient results.
3) To personalize search according to user interest.

Index Terms— **Web crawling, Indexing, QD miner**.

I. INTRODUCTION

There are different ways to help users to better describe their query information need. Query reformulation and query recommendation (or query suggestion) are two popular ways out of them. The main goal of facets mining is different from query recommendation and reformulation.

The goal of the facet mining is to summarize the knowledge and information contained in the query. A query facet is a collection of related and informative words which describes important aspects of the query. Here a facet item is typically a word. A Web query has multiple facets that summarize the information about the query from different perspectives. If the user enter 'watches' as query then different aspects of the query 'watches' are displayed as facets which includes gender categories, brands, styles, colors, etc. Facets are assigned weight in order to display the facets priority wise. Techniques used in various approach as seen in Table I.

II. METHODS USED

Following methods collectively called QD miner are used:

a) URL Extraction:

This method is used to extract the seed sites from sources like Google, yahoo, Bing, etc. When the end user enters his query the search engine sources displays hundreds of the link

with reference to the entered query. Top matching URL's containing the query words in them are extracted by smart crawling. For that, reverse searching algorithm is used.

b) Content Extraction:

This method is used to extract the web contents from extracted URL's. Document parsing is done to extract the contents. In Document parsing all the word elements in HTML tags (like select, ul, ol, Table) of the web pages are extracted. From each document we extract the set of content lists.

c) Mirror Websites Removal:

In this method two websites with the different URL's may contain the duplicated contents. It generates duplicated extracted content list. Fine grained similarity is calculated between the two lists by based on Hamming Distance between their contents. One of the duplicated list is then removed so that results are more fined and without redundancy.

d) List Weighting:

Some of the extracted facets lists can be noisy or unimportant. Good lists more frequently occur in many websites and contain the informative items. Therefore we calculate weight age of each list based on two components 1) Frequency of Occurrence 2) IDF (Inverse Document Frequency)

e) List Clustering:

In this method facet lists containing the similar items are clustered together. For this, QT (Quality Threshold) algorithm is used.

f) Item ranking and display:

In this method items are ranked according to their frequency of occurrence. Finally, highly ranked items are displayed before low rank items in descending order as "facets".

III. FINDINGS

Query facet extraction is evaluated with different perspectives:

1) Quality of clustering
2) Facet ranking effectiveness
3) Effectiveness in finding facets

Using Different metrics the all the above facet extraction perspectives are evaluated in order to get good quality facets.

Existing facet mining systems focused on to generate the summaries by using sentences extracted from the documents while QD miner system generates facets based on frequent lists. QD mining approach is different than the existing approach in two ways:

1) Open domain: Queries are not related to specific domain.
2) Query dependant: Facets are extracted from top retrieved documents for each query

Again QD mining approach uses three patterns to extract out the facet list from the web pages. The three techniques are free text pattern, HTML tag pattern, and repeat region pattern. Results shows that combination of these three patterns gives the best performance proving that QD mining approach is more efficient than the existing facets mining approaches.

IV. LITERATURE REVIEW

Automatically Mining Facets for Queries [1] from their search results-This Survey proposed the systematic solution for facets mining. Facets are extracted from the seed sites. These seed sites are the sites which we get as result when we do web search for our queries. From these top seed sites facets are extracted by document parsing, weighting, clustering and ranking of the extracted facets.

Query Subtopic Mining by combining [2] Multiple Semantics-The framework of the proposed method is divided into three parts, Aspect Phrase Extraction, Semantic Representations and Clustering & Subtopic Mining. In the first part, the related queries of the topic (original query) are extracted from the query log and denote the query with multi-word phrase. Then, novel semantic representations and combinations are used to represent the query aspect phrases for distinguishing the semantics of words, such as, the synonymous with special-shapes or words with different meanings. Finally, they adopt the clustering approach to generate the subtopics and each cluster denotes one subtopic of the initial query.

Search Result Diversification [3] based on Query Facets- In this paper researchers propose three faceted models which diversify search results based on the faceted subtopics. They again adopt the diversification algorithm which improve the result diversity.

Beyond basic faceted search-This paper [4] describes two extensions to the basic faceted search system. The extensions adds to the faceted applications by flexible and dynamic business data collection and this enable the users to gain insight into their data which is very rich quality of data because it is much more than just knowledge of the documents belonging to each facet

Dynamic faceted [5] search for discovery driven analysis-This paper implements OLAP style discovery driven analysis on big set of unstructured and structured data. Researchers again propose a new technique to measure the interestingness and novel navigation method to set the users expectation. Again it takes feedback from user and feedback survey results validate that the proposed approach meets expectations and is promising. They build the efficient run time engine on the top of the inverted index by exploiting codes and bit set tree.

Extracting Query [6] Facets from Search Results -This survey proposed new evaluation metric for this task to combine recall and precision of facet terms with grouping quality. To learn query facets experimental results shows that the supervised method classifies other unsupervised methods.

Optimal Algorithms for Crawling a Hidden Database [7] in the Web -This paper resolves the problem of relevant search for the user in order to mine out all the tuples from database by implementing some efficient algorithms which does the task to find the relevant search results even in the worst scenario by taking only small number queries as input .Researchers also propose the theoretical results which indicates that these algorithms are optimal.

A Two-stage Crawler for Efficiently Harvesting [8] Deep-Web Interfaces - This paper proposed the mechanisms in two stage crawler called smart crawler is used for efficient mining of the deep web pages .There are two steps involved to mine deep web pages. In first step this smart crawler does the site based searching for the centre pages with the help of any search engine which is able to avoid the visiting of large number of web pages .In order to achieve the more accurate and correct results for focused crawler this smart crawler gives ranking according to the priority .High priority sites are ranked top. In the second stage smart crawler does fast searching by extracting most relevant links. Researchers design the link tree data structure in order to achieve the broader coverage

Searching Documents [9] based on Relevance and Type-The paper implements the problem in a general framework consists of ‘type model’ and ‘relevance model’. The relevance model indicates whether or not a document is relevant to a query. The type model shows that whether the document does belong to the designated document type or not. Researchers consider three methods for combing the models: linear combination of scores thresh holding on the type score and hybrid of the previous two methods. It takes course page search and instruction document search as examples and they have conducted a series of the experiment.

Personalization on E-Content Retrieval Based on Semantic Web Services - This model proposes an approach [10] for filtering the educational content based on Case Based Reasoning. It is based on the model Architecture for Intelligent Recovery of the Educational content in the Heterogeneous Environment. Multi agent architecture search and integrate heterogeneous educational content through the recovery model which uses the federated search. The technologies and model which are presented in this research paper exemplify the potential for developing personalized recovery systems for digital content based on the paradigm of the virtual organizations of agents. The advantages of the architecture proposed in this paper are its flexibility, customization, and efficiency.

Facetedpedia: Dynamic Generation of Query-Dependent Faceted [11] Interfaces for Wikipedia - In this paper sharing, publishing, and connecting data on the Web provides new alternative for data integration and interoperability. However, proliferation of distributed and interconnected data sources on the Web creates significant new challenges for continuously

managing the large number of vast data sets and their inter dependencies. In these article researchers focuses on the main problem of preserving evolving structured interlinked data. They propose that a number of issues which hinder applications and users are related to the temporal aspect which is intrinsic in Linked Data. This work propose three use cases in order to motivate the approach and also discuss issues that occur and propose way to the solution

Query Recommendation using Query Logs [12] in Search Engines – This survey presents a novel query processing technique which maintains high accuracy and scalability, and again it manages to minimize the latency to great extent in answering location based spatial queries. Proposed approach depends on peer-to-peer sharing, which enables to process queries without delay at a mobile host by using query results cached in its neighboring mobile peers.

Translating Queries into Snippets [13] for Improved Query Expansion - Proposed work uses the approach of keyword mining. Indexing approach is applied over search data. Spatial inverted index extends the standard inverted index which address multidimensional information. It comes with algorithms which answer the nearest neighbor queries with keywords.

V. EXISTING SYSTEM OVERVIEW

Users need to frequently modify their search query in order to get desired result for their web queries. This strategy of query modification is called as query reformulation. Different kinds of existing systems have proposed different approaches to get the desires query results. But automatic facet Mining approach is different and most effective approach to get desires results for the users entered queries.

Existing systems used following different kinds of strategies:

Computer generated reformulations:

By using query logs new query reformulation ways has been discovered. Again by using click behaviour automatically generated reformulations were discovered

Query session boundary detection:

Session is series of interactions done by the user in order to get their desired information. Session boundary detection is done to discover different query reformulation strategies.

Click data analysis:

Click data indicates the search result preference. So click data analysis is done in order to improve search relevance.

Disadvantages:

- High computational time.
- Results with less accuracy and efficiency.

TABLE I: THE TECHNIQUES USED IN VARIOUS APPROACH

No.	Paper Name	Techniques
1)	Query Subtopic Mining by Combining Multiple Semantics	1) Clustering Query Reformulation 2) phrase embedding representation and query category distributional representation
2)	Search Result Diversification Based on Query Facets	1) Intent-aware diversification algorithms that s user intents as subtopics. 2) faceted diversification approaches
3)	Beyond basic faceted searching	1)Multifaceted search 2) On line analytical processing to efficiently and intuitively support analysis of multi-dimensional data at multiple aggregation levels
4)	Extracting Query Facets from Search Results	URL extraction, content extraction, facets clustering, facets ranking
5)	Searching Documents Based on Relevance and Type	1)Relevance Model Framework 2)Type Model Framework
6)	Facetedpedia: Dynamic Generation of Query-Dependent Faceted Interfaces for Wikipedia	faceted interface discovery algorithms that optimize the ranking metric
7)	Query Recommendation using Query Logs in Search Engines	1) Query clustering process by which groups of queries are identified which are semantically similar 2) Uses the content of historical preferences of users in the query logs.
8)	Translating Queries into Snippets for Improved Query Expansion.	1)Translation Model- based on the sequence of alignment models which contain null words 2)Language Model-Assigns probability to string of words
9)	Optimal Algorithms for Crawling a Hidden Database in the Web	1) Matching Function. 2) Label Matching.
10)	Dynamic faceted Discovery-driven analysis	1) Case based reasoning 2) Federated Search

VI. APPLICATIONS

Facet mining technique can be used for different kinds of applications. This technique is used for huge library database applications and information science research applications and to some computer science research applications and commercial search applications Eg.Amazon.com need facets mining application in order to get required data in efficient manner

VII. CONCLUSION

This survey is performed with intent to collect various facet mining techniques. Different types of facet mining mechanism are analyzed. A query facet is single word or set of words which summarizes important information about the query. Facet mining mechanism proves very useful as it saves the searching time of the user. It improves the searching experiences of the user aiding him to have all the relevant links of the websites containing most relevant information for his entered query on the same page. This facet mining technique is mostly useful for e-commerce applications, search engines, huge research library applications, etc.

ACKNOWLEDGMENT

I take this chance to express my appreciation to my guide and Head of the Department of Computer Engineering, RMDSSOE, Prof. Vina M. Lomte for her kind cooperation and guidance during the entire research work. I would also like to thank our Principal and Management for providing lab and other facilities.

REFERENCES

[1] Zhicheng Dou, Member, IEEE, Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song Automatically Mining Facets for Queries from their search results. IEEE transactions on knowledge and data engineering, vol. 28, no. 2, Feb. 2016

[2] Lizhen Liu, WenbinXu, Wei Song, HanshiWang and Chao Du. Query Subtopic Mining by Combining Multiple Semantics. International Journal of Multimedia and Ubiquitous EngineeringVol.10, No.12 (2015).

[3] Sha Hu, Zhi-Cheng Dou, Xiao-Jie Wang. Search Result Diversification Based on Query Facets. Journal of computer science and technology 30(4): 888–901 July 2015.

[4] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D.Sheinwald, E. Shekita, B. Sznajder, and S. Yogev Beyond basic faceted searching. Proc. Int. Conf. Web Search Data Mining, 2008, pp. 33–44.

[5] D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman, Dynamic faceted search for Discovery-driven analysis in ACM .Int. Conf. Inf. Knowl. Manage, pp. 3–12, 2008.

[6] Weize Kong and James Allan Extracting Query Facets from Search Results. Center for Intelligent Information Retrieval School of Computer Science University of Massachusetts Amherst, MA 01003.

[7] Cheng Sheng1 Nan Zhang3 Yufei Tao1, 2Xin Jin3, Optimal Algorithms for Crawling a Hidden Database in the Web. Istanbul, Turkey. Proceedings of the VLDB Endowment, Vol. 5, No. 11.

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

[9] Jun Xu1, Yunbo Cao1, Hang Li1, Nick Craswell2, and Yalou Huang3,Searching Documents Based on Relevance and Type in ECIR 2007, LNCS 4425, pp. 629 – 636, 2007.

[10] A.B. Gil1, S. Rodríguez1, F. de la Prieta1 and De Paz J.F., Personalization on E-Content Retrieval Based on Semantic Web Services in Department of Computer Science, University of Salamanca, Plaza de la Merced, Salamanca 37008, Spain.

[11] Chengkai Li, Ning Yan, Senjuti B. Roy, Lekhendro Lisham, Gautam Da Facetedpedia: Dynamic Generation of Query-Dependent Faceted Interfaces for Wikipedia, in WOD '12, May 25 2010, Nantes, France

[12] Ricardo Baeza-Yates1, Carlos Hurtado1, and Marcelo Mendoza, Query Recommendation using Query Logs in Search Engines, in ECIR 2007, LNCS 4425,pp. 629 636, 2009.

[13] Stefan Riezler and Yi Liu and Alexander Vasserman, Translating Queries into Snippets for Improved Query Expansion in International journal of computer science Vol. 2, Issue 2, pp: (82-99), Month: April-June 2014.

AUTHOR'S BIOGRAPHY:

Duhita Pawar
Place: Pune
Email: duhita.pawar24@gmail.com
Education: B.E (Information Technology) from Sant Gadge Baba Amravati University, Amravati., ME (Computer Engineering) from Savitribai Phule University, Pune.
Student in Department of Computer Engineering RMD, Sinhgad School of Engineering Pune, 411502, India.

Prof. Vina M. Lomte
Place: Pune
Email: vina.lomte.rmdssoe@sinhgad.edu
Education: M.E (Comp. Engg.), B.E. (CSE) Ph.D.* from Savitribai Phule University, Pune.
Professor, Department of Computer Engineering, RMD, Sinhgad School of Engineering Pune, 411502, India.