IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

A Survey on Mining Aspects for Queries

Haritha Padmanabhan[1], Derroll David[2]

PG Student, Computer Science & Engineering, Vimal Jyothi Engineering College, Kannur, India[1]

Assistant Professor, Computer Science & Engineering, Vimal Jyothi Engineering College, Kannur, India [2]

Abstract: There is a problem of finding query facets or angles that is a particular aspect or feature of something which are multiple groups of words or phrases that explain and summarize th content covered by a query. Query facets provide interesting and useful knowledge about a query and thus can be used to improve search experiences in many ways. First, we can display query facets together with the original search results in an appropriate way. Second, query facets may provide direct information or instant answers that users are seeking. Third, query facets may also be used to improve the diversity of the ten blue links. It is an assumption that the important aspects of a query are usually presented and repeated in the query's top retrieved documents in the style of lists, and query facets can be mined out by aggregating these significant lists. This can be solved by automatically mine query facets by extracting and grouping frequent lists from free text, HTML tags, and repeat regions within top search results. Then a large number of lists do exist and useful query angles can be mined.

Keywords: Facets, Mining, Query, Web page, HTML.

I. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years, due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from market analysis, fraud detection, and customer retention, to production control and science exploration. Data mining can be viewed as a result of the natural evolution of information technology.
One of the problem in quering is finding query facets or different angles of a query. A query facet is a set of items which describe and summarize one important aspect of a query. Here a facet item is typically a word or a phrase. A query may have multiple facets that summarize the information about the query from different perspectives. Facets for the query "watches" include the knowledge about watches in different unique aspects, including brands, gender categories, supporting features, styles, and colors. The query "visit Beijing" has a query facet about popular resorts in Beijing like tiananmen square, forbidden city, summer palace, etc. and a facet on travel related topics like attractions, shopping, dining, etc.

Query facets can be used to improve search experiences in many ways because they provide interesting and useful knowledge about a query. First, we can display query facets together with the original search results in an appropriate way. Thus, users can understand the important charecteristics of a query without browsing tens of pages. For example, a user could learn different brands and categories of watches. It can also implement a faceted search based on the mined query facets. User can clarify their specific need by selecting facet items. Then results of search could be restricted to the documents that are relevant to the items. A user could drill down to women's watches if he is looking for a gift for his wife. These multiple groups of query facets are particularly useful in case of vague or ambiguous queries, such as apple, windows etc. Apple could show the apple company, iphone, laptop etc. in one facet and different types of the fruit apple in another. Second, query facets may provide direct information or instant answer that users are seeking. For example, for the query related to a serial, all episode titles are shown in one facet and main actors are shown in another facet. In this case, browsing time can be saved by displaying the query facets. Third, query facets may also be used to improve the diversity of the ten blue links. Query facets also contain structured knowledge covered by the query, and thus they can be used in other fields besides traditional web search, such as semantic search or entity search [1].

Query Reformulation and Recommendation

Query reformulation and query recommendation or query suggestion are two popular ways to help users to describe their information need better. Query reformulation is the process of modifying a query that can better match according to the user's information need. The query recommendation techniques generate alternative queries semantically similar to the original query to give suggesions to the user. The main goal of mining facets is different from query recommendation. The facet mining is to summarize the knowledge and information contained in the given query. The query recommendation is to find a list of related or expanded queries. However, query facets include semantically related phrases or terms that sometimes can be used as query reformulations or query suggestions. It is different from

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

transitional query suggestions and it can utilize query facets to generate structured query suggestions, that is, multiple groups of semantically related query suggestions. This potentially provides richer information than traditional query suggestions and might help users find a better query more easily.

Query-based Summarization

Query facets are a specific type of summaries which describe the main topic of the given text. Existing summarization algorithms are classified into different categories in terms of their summary construction methods like abstractive or extractive, the number of sources for the summary like single document or multiple documents, types of information in the summary like indicative or informative, and the relationship between summary and query like generic or query-based. Facet mining aims to offer the possibility of finding the main points of multiple documents. So that save users' time on reading whole documents by clicking on each url. The difference is that most existing summarization systems dedicate themselves to generating summaries using sentences extracted from documents.

Entity Search

In recent years, the problem of entity search has received much attention. Its goal is to answer information needs that focus on entities. As for some queries, mining query facets is related to entity search and facet items are kinds of entities or attributes. Some existing entity search approaches also exploited knowledge from structure of webpages. Finding query facets differs from entity search in the some aspects. Firstly, finding query facets is applicable for all queries, rather than just entity related queries. Secondly, they tend to return different types of results. The result of an entity search are entities, their attributes, and associated homepages, whereas query facets are comprised of multiple lists of items, which are not necessarily entities.

Query Facets Mining and Faceted Search

Faceted search is a technique for allowing users to digest, analyze, and navigate through multidimensional data. It is widely applied in e-commerce and digital libraries. Most existing faceted search and facets generation systems are built on a specific domain such as product search or predefined facet categories. For example, Dakka and Ipeirotis [2] introduced an unsupervised technique for automatic extraction of facets that are useful for browsing text databases. Facet hierarchies are generated for a whole collection, instead of for a given query.Facets of a query are automatically mined from the top web search results of the query without any additional domain knowledge required. As query facets are good summaries of a query and are potentially useful for users to understand the query and help them to explore information.

They are possible data sources that enable a general open-domain faceted exploratory search.

Faceted Search

To make it easier for users to quickly find the most relevant search results, add faceting functionality. With faceting, search results are grouped under useful headings, using tags you apply ahead of time to the documents in your index. For example, the results of a shopping query for books might be grouped according to the type of book and the price. Use of faceted search is to control the presentation of search results. In the end, it's about helping the user find the right information. Faceted search gives a user the power to create an individualized navigation path, drilling down through successive refinements to reach the right document. This more effectively mirrors the intuitive thought patterns of most users. Faceted search has become an expected feature, particularly for commerce sites. Faceted search is performed in several parts:

- Index: To each document in the index, add tags to specify a value for each facet. For example, for each book in the index, tag it with the type of material and the price range.
- Search results: For every search, the Searchify server returns a count of how many matching documents were tagged with each value within each facet. For example, if the query was for "books," we might find out that in the facet "type of material," our index contains 13 science fiction books, 15 romance novels, and 10 cookbooks; and in the price facet, there are 5 books under $10, 200 books from $10-19.99, and so on.
- Query: We can include facet values as query criteria. For example, you can write a query that returns only the romance novels under $10.
- Web page: Use the facets and document counts returned by the server to create a set of facet links on your web page. Then construct queries to be activated by each facet link, passing in the appropriate values.

II. LITERATURE REVIEW

There are different works in the field of mining of facets and subtopics to improve searching results. O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev proposes a method [3] that is beyond basic faceted search. Interaction with complex and high-dimensional data

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

is often involved in transactions. Accordingly, intuitive and easy interaction modes must be featured in information discovery and e-commerce systems to allow non-experts to explore data. Multifaceted search is a popular and intuitive interaction paradigm for discovery and mining applications that allows users to digest, analyze and navigate through multidimensional data, also known as guided navigation. Faceted search applications are implemented in many web sites especially in e-commerce sites. Therearemultiple steps involved in a typical user's interaction with a faceted search interface, that is, (1) type or refine a search query, or (2) navigate through multiple, independent facet hierarchies which describe the data by drill-down (refinement) or roll-up (generalization)data mining operations. When certain values across several facets are chosen as the current context of search, faceted applications show refinements of those facets (categories), that are possible, to sub-categories, typically along with the number of search resultswhich satisfying both the freetext query and the current facet constraints, present in each subcategory.By presenting a quantitative overview on the variety of data available, these counts provide guidance to the user. So, hinting at the refinement operations that seem most promising for zooming in on the target information need. Nevertheless, By providing richer insight into the data, guided navigation interfaces can be greatly improved . The ability to view flexible and dynamic aggregations over faceted data are typically found in business intelligence applications over structured data. It would allow users to make more informed drill-down and roll-up choices, which successively help them in making better decisions. There is another shortcoming
for faceted search which is that its basic data model.In that documents are associated with sets of values across several independent facet hierarchies. It is too restrictive to model some real-life data.

M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava proposed [4] faceted search and browsing of audio content on spoken web. The "Spoken Web Search" task involves searching for an audio content within an audio content using an audio content query. This task needs researchers to build a language-independent audio search system. So that, when user given an audio query, it should be able to find the appropriate audio file or files and the approximate location of a query term within the audio file. Evaluation is performed using standard NIST metrics. As a contrastive condition, participants can submit systems not based on an audio query. The lexical form of the query cannot be used to infer the language in the audio-only condition. The goal of the task is primarily to compare the performance and limitations of different approaches on this type of task and data, not a performance comparison between different sites.

D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman proposed a dynamic faceted search for discovery-driven analysis [5]. A dynamic faceted search system for discovery driven analysis on data with both textual content and structured attributes. From a keyword query, it is needed to select a small set of "interesting" attributes dynamically and present aggregates on them to a user. Similar to work in OLAP exploration, define "interestingness" as how surprising an aggregated value is, based on a given expectation. Two new contributions by proposing a novel navigational expectation, which is particularly useful in the context of faceted search, and a novel interestingness measure through judicious application of p-values. It is an efficient dynamic faceted search system by improving a popular open source engine, Solr. Solr is the second-most popular enterprise search engine after Elasticsearch. Solr runs as a standalone full-text search server. System utilizes compressed bitmaps for caching the posting lists in an inverted index. A novel directory structure called a bitset tree for fast bit set intersection. Conduct a comprehensive experimental study on large real data sets and show that the engine performs 2 to 3 times faster than Solr.

Yunhua Hu, Yanan Qian, Hang Li, Daxin Jiang, Jian Pei,and Qinghua Zheng proposes a method [6] for mining query subtopics from search log data. Identifying the major senses and angles of queries from search log data, referred to as query subtopic mining. It is a major issue in web search. From search log analysis, it was found that there are two interesting phenomena of user behavior that can be grasped to identify query subtopics. It is referred to as 'one subtopic per search' and 'subtopic clarification by keyword'. One subtopic per search means, if a user clicks multiple URLs in one query, then the clicked URLs tend to represent the same sense or facet. Subtopic clarification by keyword means that users often add an additional keyword or keywords to expand the query for clarifying their search goal. Thus, the keywords tend to be symptomatic of the sense or facet. The proposed a clustering algorithm that can effectively grasp the two phenomena to mine the major subtopics of queries automatically, where each subtopic is represented by a cluster that containing a number of URLs and keywords. The mined subtopics of queries can be used in multiple tasks in web search. This method can significantly improve the efficiency of users' ability to find information.

Weize Kong and James Allan describes a method [7] for extracting query facets from search results. Web search queries are often ambiguous or multi-faceted, which makes a simple ranked list of results insufficient. To assist information finding for such faceted queries, explore a technique that explicitly represents interesting facets of a query using groups of semantically related terms which are extracted from search results. As an example, for the query "baggage allowance", these groups might be different airlines, different flight types like domestic, international, etc. or different travel classes like first, business, economy, etc. These groups are knownas query facets and the terms in these groups are facet terms. The developed approach is based on a graphical model to recognize query facets from the noisy candidates found. The graphicalmodel absorbes how likely a candidate term is to be a facet term as well as how likely two terms are to be grouped together in a query facet, and captures the dependencies between the two factors. Proposed two algorithms for approximate inference on the graphical model since exact inference is intractable.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

The evalation combines recall and precision of the facet terms with the grouping quality. Chengkai Li, Ning Yan, Senjuti B. Roy, Lekhendro Lisham and Gautam Das suggests a new method [8] for dynamic generation of query-dependent faceted interfaces for wikipedia. It proposes Facetedpedia, a faceted retrieval system for information discovery and inspection in Wikipedia . If a set of Wikipedia articles is given, it is resulting from a keyword query, Facetedpedia generates a faceted interface for navigating the result articles. Compared with other faceted retrieval systems, in both facet generation and hierarchy construction, Facetedpedia is fully automatic and dynamic, and the facets are based on the rich semantic information from Wikipedia . The essence of this approach is to build upon the combined vocabulary in Wikipedia, more specifically the intensive internal structures like hyperlinks and folksonomy (category system). Given the sheer size and complexity of this corpus, the space of possible choices of faceted interfaces is prohibitively large. They propose metrics for ranking individual facet hierarchies by user's navigational cost, and metrics for ranking interfaces by both their average pairwise similarities and average navigational costs.

Xu Ling, Qiaozhu Mei, ChengXiang Zhai and Bruce Schatz proposed a method [9] for mining multi-faceted overviews of arbitrary topics in a text collection. To generate a multi-faceted overview of a topic in a text collection is a common task in many text mining applications. Such an overview not only directly serves as an informative summary of the topic, but also provides a detailed view of navigation to different topic facets. Some of the existing work has describe this problem as a categorization problem and requires training examples for each facet. This has three limitations: (1) All facets are predefined, so that may not fit the need of a particular user. (2) Training examples for each facet are sometimes unavailable. (3) Approach only works for a predefined type of topics. Multi-faceted mining break these limitations. It studies a more realistic new setup of the problem, in which it would allow a user to flexibly describe each facet with keywords for an arbitrary topic and attempt to mine a multi-faceted overview in an unsupervised way. They attempt a probabilistic approach to solve this problem. They are also more informative than unstructured flat summaries. The method is quite general, thus can be applied to multiple text mining tasks in different application domains.

Jeffrey Pound, Stelios Paparizos, and Panayiotis Tsaparas proposed a query-log mining approach [10] for facet discovery for structured web search. There has been a strong trend of incorporating results from structured data sources into keyword-based web search systems such as Bing or Amazon, in recent years. Facets are a powerful tool for navigating, refining, and grouping the results, when presenting structured data. For a given structured data source, finding an ordered selection of attributes and values that will populate the facets is a fundamental problem in supporting faceted search. This creates two types of challenges. First, because of the limited screen real-estate, it is important that the top facets best match the anticipated user intent. Second, the large scale of available data to such engines demands an automated unsupervised solution. The model propose the user faceted-search behavior using the intersection of web query-logs with existing structured data. A challenge in this approach is the inherent ambiguity in mapping keywords to the different possible attributes of a given entity type, since web queries are formulated as free-text queries,. It present an automated solution that extracts user preferences on attributes and values, employing different disambiguation techniques ranging from simple keyword matching, to more sophisticated probabilistic models.

Peiling Wang and Michael W. Berry and Yiheng Yang describes a method [11] for mining longitudinal web queries and their trends and patterns. The purpose of the study is three: (1) to understand Web users' query behavior; (2) to identify problems encountered by these Web users; (3) to develop appropriate techniques for optimization of query analysis and query mining. The linguistic analyses focus on query structures, lexicon, and word associations using statistical measures such as Zipf distribution and mutual information. A data model with finest granularity is used for data storage and iterative analyses.

Raymond Chi-Wing Wong, Jian Pei, Ada Wai-Chee Fu, and Ke Wang explains a method [12] for mining favorable facets. In multi-criteria decision making applications, the importance of dominance and skyline analysis has been well recognized. Most previous studies assume a fixed order on the attributes. In practice, different customers may have different preferences on nominal attributes. So, identifing an interesting data mining problem, finding favorable facets. It has not been studied before. Given a set of points in a multidimensional space, for a specific target point p. It is to discover with respect to which combinations of orders (e.g., customer preferences) on the nominal attributes p is not dominated by any other points. Such combinations are called the favorable facets of p. It consider both the effectiveness and the efficiency of mining. There will be many favourable facets for given point. They propose the notion of minimal disqualifying condition (MDC) which is effective in summarizing favorable facets and develop efficient algorithms for favorable facet mining for different application scenarios. The first method computes favorable facets on the fly. The second method pre- computes all minimal disqualifying conditions to lookup the favourable facets in constant time.

In all of the search engines, when user search for a keyword or query, it will display a plenty of results because, a query has various aspects or facets. These are may or maynot be useful to the user. User unhappy when no result on first page satisfies information need and results misleadingly appear relevant. It is a tedious process to findout the exact results. By using the first search results of the query and attributes of the search results, the facets can be mined out. So

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

that, the user can easily find out the direction in which he should search and avoid the effort of exploring each search results.

III.CONCLUSION

Query facets provide interesting and useful knowledge about a query and thus can be used to improve search experiences in many ways. First, we can display query facets together with the original search results in an appropriate way. Second, query facets may provide direct information or instant answers that users are seeking. Third, query facets may also be used to improve the diversity of the ten blue links. There is a problem of finding query facets. It is a systematic solution, to automatically mine query facets by aggregating frequent lists from free text, HTML tags, and repeat regions within top search results.

REFERENCES

[1] Z. Dou, Z. Jiang, S. Hu, J.-R. Wen, and R. Song, "Automatically mining facets for queries from their search results," IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 2, pp. 385–397, 2016.
[2] W. Dakka and P. G. Ipeirotis, "Automatic extraction of useful facet hierarchies from text databases," in Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. IEEE, 2008, pp. 466–475.
[3] 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 search," in Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, 2008, pp. 33–44.
[4] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava, "Faceted search and browsing of audio content on spoken web," in Proceedings of the 19th ACM international conference on Information and knowledge management. ACM, 2010, pp. 1029–1038.
[5] D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman, "Dynamic faceted search for discovery-driven analysis," in Proceedings of the 17th ACM conference on Information and knowledge management. ACM, 2008, pp. 3–12.
[6] Y. Hu, Y. Qian, H. Li, D. Jiang, J. Pei, and Q. Zheng, "Mining query subtopics from search log data," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012, pp. 305–314.
[7] W. Kong and J. Allan, "Extracting query facets from search results," in Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2013, pp. 93–102.
[8] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das, "Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia," in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 651–660.
[9] X. Ling, Q. Mei, C. Zhai, and B. Schatz, "Mining multi-faceted overviews of arbitrary topics in a text collection," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2008, pp. 497–505.
[10] J. Pound, S. Paparizos, and P. Tsaparas, "Facet discovery for structured web search: a query-log mining approach," in Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. ACM, 2011, pp. 169–180.
[11] P. Wang, M. W. Berry, and Y. Yang, "Mining longitudinal web queries: Trends and patterns," Journal of the American Society for Information Science and Technology, vol. 54, no. 8, pp. 743–758, 2003.
[12] R. C.-W. Wong, J. Pei, A. W.-C. Fu, and K. Wang, "Mining favorable facets," in Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2007, pp. 804–813.
[13] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text miningan overview," International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pp. 7–16, 2015.