International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2015): 78.96 | Impact Factor (2015): 6.391

Survey on Query Facets Mining Approaches

Sheetal Sonwane[1], Nilam Patil[2]

[1]M.E. Student, Department of ME Computer Engineering, D.Y.Patil College of Engineering, Akurdi, Savitribai Phule Pune University-411044, Pune, India

[2]Asst. Prof., Department of ME Computer Engineering, D.Y.Patil College of Engineering, Akurdi, Savitribai Phule Pune University-411044, Pune, India

Abstract: ***Important features of a query are generally presented and repeated in the top retrieved documents in the style of lists. Query facets can be extracted by collecting these significant lists. Query facets may provide direct information or instant answers that users are seeking i.e. user can select a particular facet item which he found relevant to his search need. So the list format style is much more user friendly than displaying searches sentence wise. The scope of this survey is limited to get search results of a query in list format i.e. facets. Previously there has been lot of work done for retrieving more relevant data to users in order to meet their information needs thus improving performance of search engines. Search engine provide the platform for users to describe their information need more clearly by using query facets mining. Different approaches for extraction of query facets from web search results to assist information finding for queries are discussed along with similar techniques used earlier for information retrieval of queries. Query facets represents interesting facets of a query using groups of semantically related terms extracted from search results. This paper reviews techniques that represents interesting facets of a query using groups of semantically related terms extracted from search results along with approaches that are similar to query facets mining .***

Keywords: Query Facets, faceted search, multi-faceted queries, Query Subtopic, Information search

1. Introduction

Search engines currently have become the vital tools for web users to locate information. Direct access to digital information has completely changed the rules, users can jump directly to whatever they are interested in, without consulting any complex systems. The spread of digital access to information has been accompanied by sudden increase in the volume of information available about any given topic or query, to the extent that even instant access via search doesn't necessarily make finding our search easier. Many websites now provide even more refined tools to help users find information. Filters are one such tool to find information. Filters analyze a given set of content to exclude items that don't meet certain criteria. Facets extends the idea of filters. Facets provide multiple filters, one for each different aspect of the content. Because facets describes many different dimensions of the content, it also provides a structure to help users understand the content space, and give them ideas about what is available and how to search for it.

1.1 Key concepts

a) **Query Facet**: It is a set of items which describe and summarize one important aspect of a query [1]. A query facet is defined as a set of coordinate terms i.e. terms that share a semantic relationship by being grouped under a more general hypernym ("is a" relationship) [2]. For example, Facets for the query “watches”, shown in Fig. 1. cover the knowledge about watches in five unique aspects, including brands, gender categories, supporting features, styles, and colors.

b) **Facet item**: Facet item is typically a word or a phrase in a facet. A query may have multiple facets that summarize the information about the query from different perspectives [1].

c) **Faceted Search**: It is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters. A faceted classification system classifies each information element along multiple explicit dimensions, called facets.

d) **Query Subtopic**: A query subtopic is different from a query facet. The terms in a query subtopic are not restricted to be coordinate terms or have peer relationships. Query facets organize terms by grouping "sibling" terms together. For example, {news, cnn, latest news, mars curiosity news} is a valid query subtopic for the query 'mars landing', which describes the search intent of Mars landing news, but it is not a valid query facet as the terms in it are not coordinate terms. A valid query facet that describes Mars landing news could be {cnn, abc, fox} which includes different news channels.

query: watches
1. cartier, breitling, omega, citizen, tag heuer, bulova, casio, rolex, audemars piguet, seiko, accutron, movado, . . .
2. men’s, women’s, kids, unisex
3. analog, digital, chronograph, analog digital, quartz, mechanical, . . .
4. dress, casual, sport, fashion, luxury, bling, pocket, . . .
5. black, blue, white, green, red, brown, pink, orange, yellow, . . .

Figure 1.1: Facet Example

As query facets are good summaries of a query and are potentially useful for users to understand the query and help them explore information, they are possible data sources that allow a general open-domain faceted exploratory search.

Few of the research topics related to query facets mining include Query Reformulation and Recommendation, Query-Based Summarization, Entity Search, Search Results Diversification and Search Results Clustering. They are summarized in section 2. Facets Extraction according to Data

Types reviewed in section 3. Evaluation Metrics to evaluate about the search result relevance of a query are discussed in section 4.

2. Research Topics Related to Query Facets Mining

Mining query facets is related to a number of existing research topics. In this section, we briefly review them.

2.1 Query Reformulation

The reformulation process is an iterative attempt between users and search engines in getting a satisfactory set of results[4]. Query reformulation is the process of iteratively modifying a query to improve the quality of a search engine results. Reformulations are close to the previous query both syntactically, as sequences of characters or terms, and semantically, regularly involving clear naming conventions[5]. A. Herdagdelen. et. al. [5]proposed an approach to query reformulation that provide a principled framework for the combination of string similarity and corpus-based semantic association measures using generalized Levenshtein distance algorithms. An exploration of class of models viz. unsupervised, compact and efficient for query reformulation which combines the syntactic and semantic aspects are given. J. Huang et.al.[4] studied users reformulation strategies in the context of the AOL query logs and describes the human side of query reformulation. A taxonomy of query reformulation strategies is created and built a high precision rule-based classifier to detect each type of reformulation. L. Bing et al.[6] proposed a graphical model to evaluate the quality of the candidate queries generated. The model makes use of a latent topic space, which is automatically derived from the query log, to detect semantic dependency of terms in a query and dependency among topics. The graphical model is able to capture the term context in the history query by skip-bigram and n-gram language models and is able to take the users history search interests into consideration when it conducts query reformulation for different users.

2.2 Query Recommendation

Query recommendation techniques generate alternative queries semantically similar to the original query[1]. Z. Zhang and O. Nasraoui [7] combined analysis of search engine users sequential search behavior as client-side query refinement with a traditional content based similarity method to get fast query recommendations. One dimensional graph model works by accumulating many query sessions and adding up the similarity values for many same query pairs, keeping a query's most similar queries in the final clusters. K-means or a simple K-Nearest Neighbors method can be used to obtain the related query clusters. The method is adaptive and autonomous as any fixed number of clusters is not predefined. I. Szpektor et.al.[8] address the long-tail query problem by giving influence on query template. Query templates are query constructs that abstract and generalize queries. The key idea is to identify rules between templates as means for suggesting related queries. The concept of rules between query templates is introduced which can be used to conclude recommendations for rare or previously unseen queries. The query-template flow graph is introduced as an enrichment of the query-flow graph with templates. L. Li et.al.[9] proposed a **Q**uery-**U**RL **Bi**partite based query re**C**ommendation approach, called QUBiC. QUBiC utilizes the connectivity of a query-URL bipartite graph to recommend related queries and can significantly improve the accuracy and effectiveness of personalized query recommendation systems comparing with the conventional pairwise similarity based approach. QUBiC is a framework for personalized query recommendations that uses hierarchical agglomerative clustering (HAC) for ranking of similar queries.

2.3 Entity Search

Entity retrieval is the task of finding objects related to a particular information need[10]. K. Balog et. al. [10]explore the potential of combining Information Retrieval(IR) with Semantic Web(SW) technologies to improve the end-to- end performance on a specific entity search task. To get the best of both worlds, K. Balog et. al.[10] proposed to combine text- based entity models with semantic information from the Linked Open Data (LOD) cloud. Approaches to the REF task are described using IR and SW techniques and aim to find a set of entities for each topic.

Semantic class construction tries to discover the peer or sibling relationship among terms or phrases by organizing them into semantic classes. H. Zhang et.al.[11] presented an approach that employs topic modeling for semantic class construction. Given a query q, all raw semantic classes (RASCs) are retrieved containing the item to form a collection $C_R(q)$ where C_R is collection of RASCs. Latent Dirichlet Allocation (LDA) model and probabilistic Latent Semantic Indexing Model (pLSI) are used to generate semantic classes. Offline processing is performed for $C_R(q)$ that contains RASCs and store the results on disk, in order to reduce the online query processing time.

2.4 Search Results Diversification

Search result diversification has been studied as a method that deals with ambiguous or multi-faceted queries while a ranked list of documents remains the primary output feature of web search engine today. Search result diversification tries to diversify the ranked list to account for different search target or query subtopics. A weakness of search result diversification is that the query subtopics are hidden from the user, leaving him or her to guess at how the results are organized. Query facet extraction addresses this problem by explicitly presenting different facets of a queries using groups of coordinate terms. T. Sakai et.al.[21] compare the properties of existing metrics related to the points viz. queries may have multiple intents, the likelihood of each intent given a query is available and graded relevance assessments are available for each intent. Also compared a wide range of traditional and diversified IR metrics for search results diversification after adding graded relevance assessments.

2.5 Search Results Clustering

Search results clustering is a technique that organize search results by grouping them into, usually labeled, clusters by query subtopics. Instead of organizing search results in groups, there is also work done that summarizes search results or a collection of documents in a topic hierarchy. Lawrie et al. [22] used a probabilistic model for creating topical hierarchies. A graph is constructed based on conditional probabilities of words, and the topic words are found by approximately maximizing the predictive power and coverage of the vocabulary. Lawrie et.al.[23] described a method for automatically generating hierarchies from small collections of text. Then this technique is used to summarize the documents retrieved by a search engine. Lawrie et.al.[23] show that these hierarchies provide better access to the documents than a simple ranked list and that the terms in the hierarchy are better summaries of the documents.

2.6 Faceted Search and Query Facets Mining

Faceted search is a technique for accessing information organized according to a faceted classification system, allowing users to digest, analyze and navigate through multidimensional data. It is widely used in e-commerce and digital libraries[12]. A robust review of faceted search is beyond the scope of this paper. Faceted search is similar to query facet mining. Both of them use sets of coordinate terms to represent different facets of a query[2]. But most existing works for faceted search and query facets extraction are built on a specific domain or predefined categories. Various facets extraction methods corresponding to different data types are explained and reviewed in next section.

3. Facets Extraction according to Data Types

Existing automatic facets extraction methods can be divided into three categories corresponding to the different data types: unstructured, semi-structured and structured. They are explained below.

3.1 Facet Extraction of Unstructured Data

Unstructured data refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may contain data such as dates, numbers, and facts as well. In facet term extraction, the most common form of unstructured data is the natural language text, which is always ambiguous and ill-formed.

Stoica et al. [14] proposed Castanet algorithm to select facet terms based on term frequency distribution. The main idea behind the Castanet algorithm1 is to carve out a structure from the hypernym is-a relations within the WordNet lexical database. The core of this algorithm is selecting the terms having a frequency higher than a threshold as facet term candidates for subsequent processing. This algorithm can be easily implemented and extended to different domains since only term frequency is employed.

Ling et al. [15] proposed a two-stage probabilistic method to extract facet terms based on topic model. A user is allowed to flexibly describe each facet with keywords for an arbitrary topic and attempt to mine a multi-faceted overview in an unsupervised manner. Given the original keywords from a user, this method first applies a bootstrapping algorithm to the document collection to get more correlated terms. Probabilistic mixture models are applied to these expanded terms to estimate the term distribution of every facet. This is done by simultaneously fitting the topic model to the data set and restraining the model so that it is close to the specified definition from the user. The basic idea behind the processes is to guide the topic model with user-defined keywords.

Dakka and Ipeirotis [16] proposed an unsupervised automatic facet extraction algorithm using external resources viz. WordNet, Wikipedia and Google for browsing text databases. This algorithm first identifies the facet term candidates in each document by using third-party term extraction services or algorithms.. Then, each candidate is expanded with context phrases appearing in external resources by querying. This step produces the latent facet terms in the expanded term set, which do not explicitly appear in the documents. At last the term distributions in the original term set and the expanded term set compared to identify the terms that can be used to construct browsing facets. This algorithm has good flexibility and extensibility. However the quality of the extracted facets heavily depends on the quality of the external resources and term extractor.

3.2 Facet Extraction of Semi-structured Data

Semi-structured data is a form of structured data that does not match with the formal structure of data models associated with relational databases or other forms of data tables i.e. does not conform to an explicit data schema but on the other hand contains tags or other markers to separate semantically related elements. Semi-structured data lies somewhere between the structured and unstructured data. Examples of the semistructured data include HTML pages, XML pages, JSON or JavaScript Object Notation . Eg. A Word document is generally considered to be unstructured data. It is possible to can add metadata tags in the form of keywords and other metadata that represent the document content and make it easier for that document to be found when people search for those terms , the data is now semi-structured. Semi-structured data has an implicit formal structure, which can be exploited to improve the quality of facet term extraction. For example, the hyperlinks of web pages can be used to evaluate the importance of facet terms.

Li et al. [17] proposed a system named Facetedpedia, a faceted retrieval system for information discovery and exploration in Wikipedia, that utilize internal hyperlinks of Wikipedia and extract facet terms automatically. Facetedpedia consider titles of articles as facet terms, and constructs the taxonomy of articles that is hyperlinked from user keywords query results, based on the Wikipedia category system.

Oren et al. [18] proposed a facet term recognition method dedicated to the semi-structured data in semantic web. This

method was implemented by dynamically constructing a faceted navigation tree based on Resource Description Framework (RDF) graph.

Sha Hu et.al.[1] explore to automatically find query dependent facets for open-domain queries based on a general web search engine. The queries are open domain meaning that they can be extracted from webpages of a search engine instead of using particular external sources like WordNet i.e. 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 explore information, they are possible data sources that enable a general open-domain faceted exploratory search. Sha Hu et.al.[1] propose aggregating frequent lists within the top search results to mine query facets and implement a system called QDMiner. QDMiner extracts lists from free text, HTML tags, and repeat regions contained in the top search results, groups them into clusters based on the items they contain, then ranks the clusters and items based on how the lists and items appear in the top results. Quality Threshold Clustering algorithm is used for clustering in QDMiner.

3.3 Facet Extraction of Structured Data

Structured data refers to any data that resides in a fixed field within a record or file. Structured data has an explicit data model or schema and is easily organized and generally stored in relational databases and spreadsheets. For structured data, the main task of facet term extraction is to select facet terms from attributes of database. Roy et al. [19] presented a method, DynaCet - a domain independent system that provides effective minimum-effort based dynamic faceted search solutions over enterprise databases. At every step, a user is asked one or more questions about the different facet terms, and the most promising set of facet terms is identified based on the user's response. Zhao et al.[20] proposed a system named TEXplorer which selects facet terms from the attributes by measuring the relevance between keywords and documents. TEXplorer can be implemented within a multi-dimensional text database, where each row is associated with structural dimensions i.e. attributes and text data Eg. a document.

4. Evaluation Metrics

Relevance metrics are used to evaluate how relevant a search result is about a given query. In facet extraction, the matching between data items and facet terms in many cases are predetermined. Objective metrics evaluate search results and search process by adopting standard benchmark. Relevance metrics is a type of objective metrics used to evaluate how relevant a search result is regarding a given query. A series of metrics have been proposed by information retrieval community to describe the binary relevance and graded relevance. Binary relevance metrics include precision, recall, F-measure, E-measure and graded relevance metrics includes mainly nDCG measure i.e. Normalized Discounted Cumulative Gain.

4.1 Normalized Discounted Cumulative Gain (nDCG)

nDCG measure is Normalized Discounted Cumulative Gain. Discounted cumulative gain (DCG) is a measure of ranking quality. In information retrieval, it is often used to measure effectiveness of web search engine algorithms or related applications. Search result lists vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of p should be normalized across queries. This is done by sorting all relevant documents in the corpus by their relative relevance, producing the maximum possible DCG through position p, also called Ideal DCG (IDCG) through that position. For yop p facets, the normalized discounted cumulative gain, $nDCG_p$, is computed as division of DCG_p and $IDCG_p$. Z. Dou et.al.[1] used nDCG evaluation metrics to measure ranking effectiveness of facets i.e. rank good facets before bad facets when multiple facets are found. Pound et al. [24] make use of nDCG to rank the output of their facet discovery algorithm.

4.2 F measure

F measure is the measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score. p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F measure can be interpreted as a weighted average of the precision and recall, where an its score reaches its best value at 1 and worst at 0. The F measure is often used in the field of information retrieval for measuring search, document classification, and query classification performance Z. Dou et.al.[1] used F measure to evaluate the quality of clusters. Xing et al. [25] used micro-F1, macro-precision, macro-recall, macro-F1 to evaluate the results of their Deep Classifier faceted search system.

5. Conclusion

In this paper, we first reviewed the research topics related to representative facets mining task then existing automatic facets extraction methods are explained. Also, evaluation metrics used for measuring quality and ranking of query facets are discussed.

Facets extraction has witnessed a booming interest recently and has been undergoing research for the past few years. Facets extraction enables users to select facets and facet terms to refine the search results in an iterative way. Focusing Query Facets Mining, the future work possible in this area includes the following:

- Part-of-speech information can be used to check the homogeneity of lists and improve the quality of query facets.
- Automatically generate meaningful descriptions of query facets.

- Semi-supervised bootstrapping list extraction algorithms can be used to iteratively extract more lists from the top results.

References

[1] Sha Hu, Ji-Rong Wen, Zhicheng Dou, Zhengbao Jiang and R. Song, "Automatically mining facets for queries from their search results," IEEE Transactions on knowledge and data engineering, pp. 385-397, 2016.
[2] W. Kong and J. Allan, "Extracting query facets from search results," in Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 93–102.
[3] D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman. "Dynamic faceted search for discovery-driven analysis," In Proceedings of CIKM '08, pages 3–12, 2008.
[4] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," in Proc. 18th ACM Conf. Inf. Knowl. Manage., pp. 77-86, 2009.
[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] L. Bing, W. Lam, T.-L. Wong, and S. Jameel, "Web query reformulation via joint modeling of latent topic dependency and term context," ACM Trans. Inf. Syst., vol. 33, no. 2, pp. 6:1–6:38, eb. 2015.
[7] Z. Zhang and O. Nasraoui, "Mining search engine query logs for query recommendation," in Proc. 15th Int. Conf. World Wide Web, pp. 1039-1040, 2006.
[8] A. G. I. Szpektor and Y. Maarek, "Improving recommendation for long-tail queries via templates," in Proc. 20th Int. Conf. World Wide Web, pp. 47-56, 2011.
[9] L. Li, L. Zhong, Z. Yang, and M. Kitsuregawa, "Qubic: An adaptive approach to query-based recommendation," J. Intell. Inf. Syst., vol. 40, no. 3, pp. 555–587, Jun. 2013.
[10] K. Balog, E. Meij, and M. de Rijke, "Entity search: Building bridges between two worlds," in Proc. 3rd Int. Semantic Search Workshop,2010, pp. 9:1–9:5.
[11] H. Zhang, M. Zhu, S. Shi, and J.-R. Wen, "Employing topic models for pattern-based semantic class discovery," in Proc. Joint Conf. 47th Annu. Meet. ACL 4th Int. Joint Conf. Natural Lang. Process. AFNLP, 2009, pp. 459–467.
[12] D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman. "Dynamic faceted search for discovery-driven analysis," In Proceedings of CIKM '08,pp. 3-12, 2008.
[13] 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., 2010, pp. 1029–1038.
[14] Stoica, E., Hearst, M.A., and Richardson, M., "Automating Creation of Hierarchical Faceted Metadata Structures," Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 2007: p.244-251.
[15] Xu Ling, Qiaozhu Mei, ChengXiang Zhai, Bruce Schatz, "Mining multi-faceted overviews of arbitrary topics in a text collection," in Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008, ACM: Las Vegas, Nevada, USA. p. 497-505.
[16] Dakka, W. and Ipeirotis, P.G., "Automatic extraction of useful facet hierarchies from text databases," in 2008 IEEE 24th International Conference on Data Engineering. 2008. p. 466-475.
[17] Chengkai Li, Ning Yan, Senjuti B. Roy, Lekhendro Lisham, Gautam Das, "Facetedpedia: dynamic generation of query-dependent faceted interfaces for wikipedia," in Proceedings of the 19th international conference on World Wide Web. 2010, ACM: Raleigh, North Carolina, USA. p. 651-660.
[18] Oren, E., Delbru, R., and Decker, S., "Extending faceted navigation for RDF data," in Proceedings of the 5th International Semantic Web Conference (ISWC). 2006. p. 559-572.
[19] S. B. Roy, H. Wang, U. Nambiar , G. Das and M. Mohania," DynaCet: Building Dynamic Faceted Search Systems over Databases," IEEE 25th International Conference on Data Engineering(ICDE), Vols 1-3. 2009. p. 1463-1466.
[20] Bo Zhao, Xide Lin, Bolin Ding, Jiawei Han, "TEXplorer: keyword-based object search and exploration in multidimensional text databases," in Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM). 2011, ACM: Glasgow, Scotland, UK. p. 1709-1718.
[21] T. Sakai and R. Song. "Evaluating diversified search results using per-intent graded relevance." In Proceedings of SIGIR '11, pages 1043-1052. ACM, 2011.
[22] D. Lawrie, W. B. Croft, and A. Rosenberg. "Finding topic words for hierarchical summarization," In Proceedings of SIGIR '01, pages 349-357, 2001.
[23] D. J. Lawrie and W. B. Croft. "Generating hierarchical summaries for web searches," In Proceedings of SIGIR '03, pages 457-458, 2003.
[24] Pound, J., Paparizos, S., and Tsaparas, P., "Facet discovery for structured web search: a query-log mining approach," in Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. 2011, ACM: Athens, Greece. p. 169-180.
[25] Dikan Xing, Gui-Rong Xue, Qiang Yang, Yong Yu, "Deep classifier: automatically categorizing search results into large-scale hierarchies," in Proceedings of the international conference on Web search and web data mining. 2008, ACM: Palo Alto, California, USA. p. 139-148.