International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

Automatically Mining Facets for Queries from Theirresules

N.Gowthami & K.Bhuvaneswari
[1]M.SC. (Computer Science), RIIMS, Tirupati,
[2]M.C.A, Department of Computer Science, RIIMS, Tirupati,
Email Id:gowthami361995@gmail.com, Email: bhuvaneswari.kagithala@gmail.com

Abstract –*In This Paper manage the issue of finding question aspects which are a few gatherings of words or expressions that clarify and audit the substance encased by an inquiry. We trust that the huge parts of a question are normally displayed and repeated in the inquiry's pinnacle recovered records in the style of records, and inquiry features can be mined out by accumulating these vital lists.propose a sorted out answer, which we allude to as QDMiner, to naturally supply question aspects by separating and gathering intermittent records from free content, HTML labels, and copy locales inside top list items. Trial result demonstrate that a major number of records are available and significant question aspects can be mined by QDMiner. We additionally break down the issue of rundown duplication, and discover unrivaled question aspects can be mined by displaying fine-grained likenesses amongst records and rebuffing the copied records.*

Keywords-Query facet, faceted search, summarize

I. INTRODUCTION

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. For example facets for the query"watches" cover the knowledge about watches in five unique aspects, including brands, gender categories, supporting features, styles, and colors. Query facets provide interesting and useful knowledge about a query and thus can be used to improve search experiences in many ways. In this work, we attempt to extract query facets from web search results to assist information finding for these queries. We define a query facet as a set of coordinate terms { i.e., terms that share a semantic relationship by being grouped under a more general a "relationship". First, we can display query facets together with the original search results in an appropriate way Thus, users can understand some important aspects of a query without browsing tens of pages. For example, a user could learn different brands and categories of watches. We can also implement a faceted search [1], [2], [3] based on the mined query facets. Second, query facets may provide direct information or instant answers that users are seeking. For example, for the query "lost season", all episode titles are shown in one facet and main actors are shown in another. Inthis case, displaying query facets could save browsing time. Third query facets may also be used to improve the diversity of the ten blue links. We can re-rank search results to avoid showing the pages that are near-duplicated in query facets at the top. 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.

Table 1
Example Query Facets Mined by QD Miner

Query: watches
1. Cartier, breitling, omega, citizen, tag heuer, bulova, casio,rolex, audemarspiguet, 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, . . .

Query: lost
1. season 1, season 6, season 2, season 3, season 4, season 5
2. matthew fox, naveenandrews, evangelinelilly, josh holloway,jorgegarcia, danieldaekim, michaelemerson
3. jack, kate, locke, sawyer, claire, sayid, hurley, desmond,boone, charlie, ben, juliet, sun, jin, . . .
4. what they died for, across the sea, what kate does, the candidate, the last recruit, everybody loves hugo, the end, . . .

Query: lost season 5
1. because you left, the lie, follow the leader, jughead, 316, . . .
2. jack, kate, hurley, sawyer, sayid, ben, juliet, locke, miles,desmond, charlotte, various, sun, none, richard, daniel, . . .
3. matthew fox, naveenandrews, evangelinelilly, jorgegarcia,henryiancusick, josh holloway, michaelemerson, . . .
4. season 1, season 3, season 2, season 6, season 4

Query: what is the fastest animal in the world
1. cheetah, pronghorn antelope, lion, thomson's gazelle, wildebeest, cape hunting dog, elk, coyote, quarter horse, . . .
2. birds, fish, mammals, animals, reptiles
3. science, technology, entertainment, nature, sports, lifestyle,travel, gaming, world business

Query: visit beijing
1. tiananmen square, forbidden city, summer palace, great wall, temple of heaven, beihai park, hutong, . . .
2. attractions, shopping, dining, nightlife, tours, tip, . . .

International Journal of Research
Available at https://edupediapublications.org/journals

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

websites are using different domain names but they are publishing duplicated content and contain the same lists. Some content initially produced by a website might be re-published by other websites, hence the same lists contained in the content might appear various times in different websites. Furthermore, different websites may publish content using the similar software and the software may generate duplicated lists in different websites.

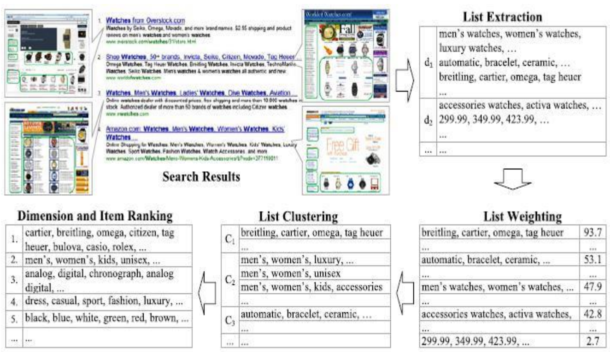

Fig. 1- System overview of QDMiner

Review of Existing Work

This section reviews the main existing work found in the scientific literature that applies on Automatically MiningFacets for Queries from Their Search Results.

[1] This paper extends established faceted search to support more affluent information discovery tasks over more difficult data models. Our first extension adds elastic, active business intelligence aggregations to the faceted application, enabling users to gain insight into their data that is far richer than just knowing the quantities of documents belonging to each facet. We see this potential as a step toward bringing OLAP capabilities, traditionally supported by databases over relational data, to the domain of free-text queries over metadata-rich content. Our second addition shows how one can proficiently extend a faceted search engine to support interrelated facets - a more intricate information model in which the values associated with a document across multiple facets are not independent. We show that by reducing the difficulty to a recently solved tree-indexing scenario, facts with correlated facets can be efficiently indexed and retrieved.

[2]Spoken Web is a network of Voice Sites that can be accessed by a phone. The substance in a Voice Site is audio. Therefore Spoken Web provides an alternate to the World Wide Web (www) in rising regions where low Internet access and low literacy are barriers to accessing the conservative www. Searching of audio content in Spoken Web through an audio query-result interface presents two key challenges: indexing of audio content is not precise, and the arrangement of results in audio is sequential, and therefore cumbersome. In this paper, we apply the concepts of faceted search and browsing to the Spoken Web search problem. We use the concepts of facets to index the meta-data associated with the audio content. We provide a means to rank the facets based on the search results. We develop an interactive query
interface that enables effortless browsing of search results through the top ranked facets. To our understanding, this is the first system to use the concepts of facets in audio search, and the first result that provides an audio search for the rural population. We present

quantitative results to illustrate the accuracy and usefulness of the faceted search and qualitative results to highlight the usability of the interactive browsing system. The experiments have been conducted on more than 4000audio documents composed from a live Spoken Web Voice Site and evaluations were carried out with 40 farmers who are the intended users of the VoiceSite.

[3]We recommend a dynamic faceted search structure for discovery- driven analysis on data with both textual content and structured attributes. From a keyword query, we want to dynamically choose a little set of ―appealing‖ attributes and present aggregates on them to a user. Similar to work in OLAP discovery, we define ―interestingness‖ as how astonishing an aggregated value is, based on a given expectation. We make two new contributions by proposing a novel―navigational‖ expectation that's chiefly helpful in the background of faceted search, and a novel interestingness measure through sensible application of p-values. Through a user survey, we find the new expectation and interestingness metric quite valuable. We develop an efficient dynamic faceted search system by improving a accepted open source engine, Solr. Our system exploits compressed bitmaps for caching the posting lists in an inverted index, and a novel directory structure called a bitset tree for fast bitset intersection. We conduct a broad experimental study on huge real data sets and show that our engine performs 2 to 3 times quicker than Solr.

[4]Faceted search helps users by presenting drill-down options as a complement to the keyword input box, and it has been used fruitfully for many vertical applications, including e- commerce and digital libraries. However, this scheme is not well explored for general web search, even though it holds great potential for supporting multi-faceted queries and exploratory search. In this paper, we discover this potential by extending faceted search into the open-domain web setting, which we call Faceted Web Search. To tackle the diverse nature of the web, we propose to use query-dependent automatic facet generation, which generates facets for a query instead of the entire corpus. To incorporate user feedback on these query facets into document ranking, we examine both Boolean filtering and soft ranking models.

[5] As the Web has evolved into a data-rich repository, with the typical ―page view,‖ current search engines are more and more inadequate. While we regularly search for a variety of data units, nowadays engines only get us in a roundabout way to pages. Hence, we propose the representation of *entity search*, a significant departure from conventional document retrieval. Towards our goal of supporting entity search, in the *WISDM*1 project at UIUC we build and assess our prototype search engine over a 2TB Web corpus. Our demonstration shows the viability and assurance of a large-scale system architecture to sustain entity search.

[6] We reflect on the task of entity search and inspect to which degree state-of-

International Journal of Research
Available at https://edupediapublications.org/journals
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 05 Issue 12
April 2018

art information retrieval (IR) and semanticweb (SW) technologies are skilled of answering information needs that focus on entities. We also investigate the potential of combining IR with SW technologies to develop the end-to- end performance on a specific entity search task. We arrive at and encourage a proposal to unite text-based entity models with semantic information from the Linked Open Data cloud.

[7]Associated entity finding is the task of returning a ranked list of homepages of significant entities of a specified type that need to engage in a given association with a given source entity. We propose a framework for addressing this task and execute a detailed scrutiny of four core components; co-occurrence models, type filtering, context modeling and homepage finding. Our initial spotlight is on recall. We examine the performance of a model that only uses co-occurrence statistics. While it identifies a set of related entities, it fails to rank them successfully. Two types of fault emerge: (1) entities of the incorrect type spoil the ranking and (2) while somehow linked to the source entity, some retrieved entities do not engage in the right relation with it. To address (1), we add type filtering based on category information obtainable in Wikipedia. To correct for (2), we add related information, represented as language models derived from documents

IV. MODELS USED

1. Unique Website Model

In the Unique Website Model, we assume that lists from the same website might contain duplicated information, whereas different websites are independent and each can contribute a separated vote for weighting facets. However, we find that sometimes two lists can be duplicated, even if they are from different websites [4], [5]. For example, mirror websites are using different domain names but they are publishing duplicated content and contain the same lists. Some content originally created by a website might be republished by other websites, hence the same lists contained in the content might appear multiple times in different websites. Furthermore, different websites may publish content using the same software and the software may generate duplicated lists in different websites. Ranking facets solely based on unique websites their lists appear in is not convincing in these cases.

2. Context Similarity Model

Hence we propose the Context Similarity Model, in which we model the fine-grained similarity between each pair of lists. More specifically, we estimate the degree of duplication between two lists based on their contexts and penalize facets containing lists with high duplication

V. QD MINER

It is observe that important pieces of information about a query are usually presented in list styles and repeated many times among top retrieved documents. So we propose aggregating frequent lists within the top search results to mine query facets and implement a system called QDMiner**.** It discovers query facets by aggregating frequent lists within the top results.

We propose this method because:

(1) Important information is usually organized in list formats by websites. They may repeatedly occur in a sentence that is separated by commas, or be placed side by side in a well-formatted structure (e.g., a table). This is caused by the conventions of webpage design. Listing is a graceful way to show parallel knowledge or items and is thus frequently used by webmasters.

(2) Important lists are commonly supported by relevant websites and they repeat in the top search results, whereas unimportant lists just infrequently appear in results. This makes it possible to distinguish good lists from bad ones, and to further rank facets in terms of importance. It automatically mine query facets by aggregating frequent lists from free text, HTML tags, and repeat regions within top search results

VI. CONCLUSION

In this paper, learn the problem of finding query facets. We propose a methodical key, which we refer to as QDMiner, to involuntarily mine query facets by aggregating recurrent lists from free text, HTML tags, and repeat regions inside top search results. We generatetwo human annotated data sets and pertain existing metrics and two new joint metrics to evaluate the superiority of query facets. Experimental results show that helpful query facets are mined by the approach. We further scrutinize the problem of duplicated lists, and find that facets can be enhanced by modeling fine-grained similarities among lists within a facet by comparing their similarities.

As the first approach of finding query facets, QDMiner can be bettered in many aspects. For example, some semi supervised bootstrapping list extraction algorithms can be used to repeatedly extract more lists from the top results. Specific website wrappers can also be engaged to extract high-quality lists from reliable websites. Adding these lists may develop both accuracy and recall of query facets. Part-of-speech information can be used to further ensure the homogeneity of lists and improve the quality of query facets. We will discover these topics to purify facets in the future. We will also inspect some other associated topics to finding query facets. Superior descriptions of query facets maybe helpful for users to improved understand the facets. Automatically produce meaningful descriptions is an fascinating research topic.

REFERENCES

[1] 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 Proc. Int. Conf. Web Search Data Mining, 2008, pp. 33–44.

[2] 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.

[3] 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.

[4] W. Kong and J. Allan, ―Extending faceted search to the general web,‖ in Proc.ACMInt. Conf. Inf. Knowl. Manage., 2014, pp. 839–848.

[5] T. Cheng, X. Yan, and K. C.-C. Chang, ―Supporting entity search: A large-scale prototype search engine,‖ in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2007, pp. 1144–1146.

[6] 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.

[7] M. Bron, K. Balog, and M. de Rijke, ―Ranking related entities: Components and analyses,‖ in Proc. ACM Int. Conf. Inf. Knowl. Manage., 2010, pp. 1079–1088.

[8] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das, ―Facetedpedia: Dynamic generation of query-dependent faceted interfaces for wikipedia,‖ in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 651–660.

[9] W. Dakka and P. G. Ipeirotis, ―Automatic extraction of useful Facet hierarchies from text databases,‖ in Proc. IEEE 24th Int. Conf. Data Eng., 2008, pp. 466–475.

[10] A. Herdagdelen, M. Ciaramita, D. Mahler, M. Holmqvist, K. Hall, models of query reformulation,‖ in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. retrieval, 2010, pp. 283–290.

[11] M. Mitra, A. Singhal, and C. Buckley, ―Improving automatic query expansion,‖ in Proc. 21st Annu. Int. ACM SIGIR Conf. Res.Develop. Inf. Retrieval, 1998, pp. 206–214.

[12] P .Anick,―Using terminological feedback for web search refinement:A log-based study,‖ in Proc. 26th Annu. Int. ACM SIGIRConf. Res. Develop. Inf. Retrieval, 2003, pp. 88–95.

[13] S. Riezler, Y. Liu, and A. Vasserman, ―Translating queries into Comput. Ling., 2008, pp. 737–744.

[14] 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.

[15] R.Baeza-Yates, C.Hurtado, and M. Mendoza, ―Query Recommendation using query logs in search engines,‖ in Proc. Int. Conf. Current Trends Database Technol., 2004, pp. 588–596.

[16] Z. Zhang and O. Nasraoui, ―Mining search engine query logs for 2006, pp. 1039–1040.