International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

Modeling reformulation queries into snippets for improving automatic query expansion

C. Ramyasree[1], Dr. A. Sureshbabu[2]
M.Tech, Dept of CSE, JNTUA, Ananthapuramu[1].
M.Tech., Ph.D, Associate Professor, Dept of CSE, JNTUA, Ananthapuramu.[2]

Abstract: Extracting query subtopics are similar to generating facets for queries. A query subtopic is often defined, as distinct information need relevant to the original query. We deal with the problem of discovering query facets, which are several groups of words or phrases that make clear, and review the content enclosed by a query. We believe that the significant aspects of a query are usually presented and recurred in the query's peak retrieved documents in the style of lists, and query facets can be mined out by aggregating these important lists. We propose an organized answer, which we refer to as QDMiner, to automatically supply query facets by extracting and grouping recurrent lists from free text, HTML tags, and duplicate regions within top search results. Experimental outcome show that a big number of lists are present and QDMiner can mine valuable query facets. We further analyze the problem of list duplication, and find modeling fine-grained similarities between lists and punishing the duplicated lists can mine superior query facets.

Keywords- Query facet, faceted search, summarization, user intent.

I. INTRODUCTION

Query facet is derived by analyzing the text query .It allows the users to explore collection of information by applying multiple filters. Faceted search / Faceted navigation is a technique for accessing information organized according to a faceted classification system. Query facets provide interesting and useful knowledge about a query. It improves search experiences. Query facet generates significant aspects from a large list of queries based on a particular product/ services. Facets access a recommendation for searched users. Automatically mine query facets that exhibit the characteristics of product/ service. A query may have multiple facets that summarize information from a query from different perspectives. We recommend a methodical solution, which we refer to as QDMiner, to automatically mine query facets by extracting and combination frequent lists from free text, HTML tags, and replicate regions within top search results. The lists can be grouped into clusters based on the items they contain, then ranks the clusters and items based on how the list and items appear in the top results. WQT (Quality Threshold with Weighted data points) is the clustering algorithm used to group high quality weighted list first. To rank query facets two models are to be used, the Unique Website Model and the Context Similarity Model. Purpose of this project is to improve the search experiences for the user and build a system that automatically mines query facets for open domain queries easily.

II. RELATED WORK

A. Extending faceted search to the general web, W. Kong and J. Allan

In this paper, we proposed Faceted Web Search, an extension of faceted search to the general Web. We studied different facet generation and facet feedback models based on our proposed extrinsic evaluation, which directly measures the utility in search instead of comparing system/annotator facets as in intrinsic evaluation. Extracting query subtopics (or aspects) is similar to generating facets for queries. A query subtopic is often defined, as distinct information need relevant to the original query. Our experiments show, by using facet feedback from users, Faceted Web Search is able to assist the search task and significantly improve ranking performance. Comparing intrinsic evaluation and extrinsic evaluation on different facet generation models, we find that the intrinsic evaluation does not always reflect system utility in real application. Comparing different facet feedback models, we find that the Boolean filtering models, which are widely used in conventional faceted search, are too strict in Faceted Web Search, and less effective than soft ranking models.

B. Beyond basic faceted search, O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. Ofek-Koifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev

Traditional faceted search to support richer information discovery tasks over more complex data models. Our first extension adds flexible, dynamic 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 capability 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 extension shows how one can efficiently extend a faceted search engine to support correlated facets a more complex information model in which the values associated with a document across multiple facets are not independent. By reducing the problem to a recently solved tree-indexing scenario, data with correlated facets can be efficiently indexed and retrieved.

C. Automatic extraction of useful Facet hierarchies from text databases, W. Dakka and P. G. Ipeirotis.

Faceted interfaces uses different orthogonal classification schemes to present in a database has been increasing worldwide. Many websites (such as YouTube, The New York Times, eBay, and Google Base)function on top of large database and offer a large type of services. In this paper, we represent an unsupervised method for automatic extraction of facets useful for browsing text databases. In order to support such exploratory interactions, the majority of the websites mentioned above use a form of concept hierarchies to support browsing on top of large sets of items. Commonly, a single hierarchy or a taxonomy that organizes thematically the contents of the database supports browsing. Unfortunately, a single hierarchy can very rarely organize the contents of database. This paper work on the automatic construction of multifaceted interfaces contributes to the development of databases. The distributional analysis step of our technique automatically identifies which concepts are important for the underlying database and generates the appropriate facet terms. Finally, we compare the term distributions in the original database and the expanded database to identify the terms that can be used to construct browsing facets. Our extensive user studies, using the Amazon Mechanical Turk service, show that our techniques produce facet switch high precision and recall that are superior to existing approaches and help users locate interesting items faster. We plan to perform more experiments in this direction and examine the performance of our techniques for a larger variety of text databases and external resources.

D. Entity search: Building Bridges between two worlds, K. Balog, E. Meij, and M. de Rijke.

The NIH was faced with the congressional imperative to create

OAM.It had formidable challenge of defining the research agenda for the new office. The OAM is brainchild of senator Tom Harkin. HE was a powerful position to influence research directions of the NIH.The United State Senate establish the Office of Alternative medicine at the national Institute of Health to facilitates evalution of medical practices.

The OAM currently functions in different capacities:

As a “bridge” between alternative and orthodox medical communities.

As a support for research through grant programs for evalution of alternative medical practices.

As Technical resource to alternative medical researches to learn good methods.

As an information resource on alternative medicine for biomedical research

Techniques for health promotion and disease prevention are integral part of medicine as an extension of primary care. The issue of the cost is an important one in the quality of life and in the quality of care option in the alternative medicine should be explode as a potential mechanism for the burden for chronic care cost.

III. PROPOSED SYSTEM ARCHITECTURE

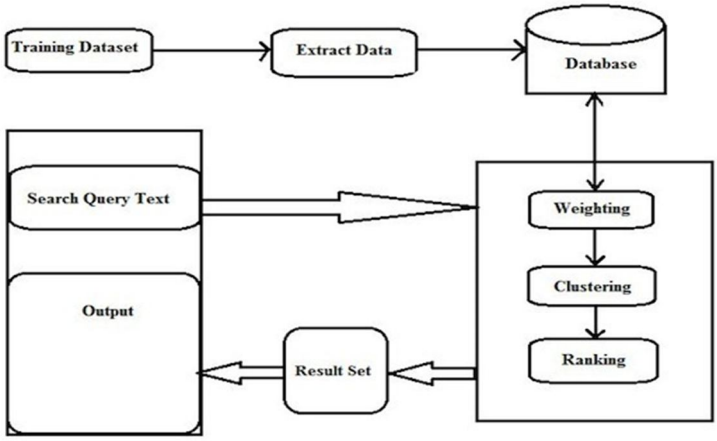

Fig: System overview of QDMiner

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

IV. MODULES DESCRIPTION

A. *List and context extraction*

From each document d in the search result set R, we extract a set of lists L_d from the HTML content of d based on three different types of patterns, namely free text patterns, HTML tag patterns, and repeat region patterns. For each extract list, we extract its container node together with the previous and next sibling of the container node as its context. We define that a container node of a list is the lowest common ancestor of the nodes containing the items in the list.

1) *Free Text Patterns - $Text_S$ and $Text_P$:* It extracts all text within document d and split it into sentences. We then employ the pattern item {, item}* (and/or) {other} item, which is similar to extract matched items from each sentence. Example: shopping for gorgeous watches from Seiko, Lucien Piccard, Citizen, and Cartier. We further use the pattern {^item (: j-) .+$}+ to extract lists from some semi-structured paragraphs. It extracts lists from continuous lines that are comprised of two parts separated by a dash or a colon. The first parts of these lines are extracted as a list.
2) *HTML Tag Patterns-HTML TAG*: To extracts lists from several list-style HTML tags, including SELECT, UL, OL, and TABLE. We named these simple HTML tag based patterns as HTMLTAG.
   a) SELECT For the SELECT tag; simply extract all text from their child tags (OPTION) to create a list. Moreover, it removes the first item if it starts with some predefined text, such as “select” or “choose”.
   b) UL/OL For these two tags, it also simply extracts text within their child tags (LI).
   c) TABLE extracts one list from each column or each row. For a table containing m rows and n columns. For each column, the cells within THEAD or TFOOT tags are regarded as table headers and are dropped from the list.

3) *Repeat Region Patterns:* The peer information is sometimes organized in well-structured visual blocks in webpages. To extract these lists, we first detect repeat regions in webpages based on vision-based DOM trees. For a list extracted from a repeat region, we choose the lowest common ancestor element of all blocks of the repeat region as a container node (i.e., the smallest element containing the entire repeat region). Note that the blocks contained in a repeat region can be non adjacent, hence the container node may not be the parent element of a block.

B. *List weighting*

To aggregate all lists of a query, and evaluate the importance of each unique list l by the following components:

1) S_{doc}: *document matching weight*: Items of a good list should frequently occur in highly ranked lists.
2) S^m: is the percentage of items contained in d
3) s^r_d measures the importance of document d.
4) S_{idf}: average invert document frequency (IDF) of item

Its most items just appear in one document in top results.

C. *List clustering*

Two lists can be grouped together if they share enough items. It use a modified QT (Quality Threshold) clustering algorithm to group similar lists. QT is a clustering algorithm that group’s data into high quality clusters. Com- pared to other clustering algorithms, QT ensures quality by finding large clusters whose diameters do not exceed a user-defined diameter threshold. This method prevents dis- similar data from being forced under the same cluster and ensures good quality of clusters.

Quality threshold algorithm (qt) Quality Threshold with Weighted data points is described as follows.

Choose a maximum diameter and minimum weight for the cluster.

Build a candidate cluster for the most important point by iteratively including the point that is closest to the Group, until the diameter of the cluster surpasses the threshold Dia max.

Here the most important point is the list, which has the highest weight. Save the candidate cluster if the total weight of its is cluster is not smaller than user defined threshold.

Recurs with the reduced set of points. In this paper, the weight of a cluster is computed based on the number of websites from which its lists are extracted.

D. Facet and item ranking

The data are ranked before it shows output this ranking occurs based on the details available in the database, for example it ranks watches as gender wise or famous brand wise or most sold wise. The top search results are extracted into related document search results and users frequently are done in the significant work in the rank based so it might easily get together by a user.

E. Modules used

1) *Unique Website Model*: In the Unique Website Model, the lists from the same website might contain duplicated information. Different websites are independent having separated vote for weighting facets. Sometimes two lists can be duplicated, even if they are from different websites. For example, mirror websites are using different domain names but they are publishing duplicated content and using 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. Some times 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

Fig: - An example of copied pages

V. CONCLUSION

Query facet is a systematic solution to automatically mine query facets by extracting and grouping frequent lists from free text. Facet based mining will help to find the attributes of a product which are prominent Facet may eliminate multi linking and multi page search method on e- commerce application. We developed a supervised method based on a graphical model to recognize query facets from the noisy facet candidate lists extracted from the top ranked search results. List extraction algorithms can be used to iteratively extract more lists from the top results. The high quality lists can be used to generate meaningful facets. These facets are generated based on the user's interest. User can navigate to the specified page by selecting the item on the facet to get detailed information. Specific website wrappers can also be employed to extract high-quality lists from authoritative websites. Good descriptions of query facets may be helpful for users to better understand the facets.

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] I. Szpektor, A. Gionis, and Y. Maarek, "Improving recommenda- tion for long-tail queries via templates," in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 47–56.
[6] K. Latha, K. R. Veni, and R. Rajaram, "Afgf: An automatic facet generation framework for document retrieval," in Proc.Int. Conf. Adv. Comput. Eng., 2010, pp. 110–114.
[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.