Challenges in QDMiner

Vidya Dhere[1], Shrikant R. Tandle[2]
[1] M. E. Student, [2] Head, Computer Science Engineering Department
[1]Computer Science Engineering Department, M.S. Bidve Engineering College, Latur, India.

***Abstract*-Query Faceted search is a way for searching users to find, analyze, and navigate through search data form online web pages. It is widely used in e-commerce and digital libraries. An effective approach for faceted search is the scope of this implementation. Most existing faceted search and facets generation systems are built on a specific domain (such as product search) or predefined facet categories. For example, Web search mining for an unsupervised contents by automatic extraction of facets that are relevant for search result for personal web search as user search interest pattern from text databases. Facet hierarchies are generated for a whole collection, instead of for a given query. Proposed facets searching system for information discovery and media exploration in online search results. Proposed system extracts and aggregates the useful semantic information from the specific knowledge database Wikipedia. In this paper, a proposed system explores to automatically find query related aspect of search for open-domain queries in Web search engine. 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.**

***Index Terms* -Query facet, faceted search, summarization, user intent**

INTRODUCTION

We deal with the problem of finding aspects of the query. A query facet is a set of elements that describe and summarize an important aspect of a query. Here a facet element is typically a word or a phrase. A query can have multiple aspects summarize the information about the different query perspectives the facets for the "watches"" query relate to knowledge observe in five unique aspects, including brands, gender categories, support functions, styles and colors. Query Faceted search is a way for searching users to find, analyze, and navigate through search data form online web pages. It is widely used in e-commerce and digital libraries. An effective approach for faceted search is the scope of this implementation. Most existing faceted search and facets generation systems are built on a specific domain (such as product search) or predefined facet categories. For example, Web search mining for an unsupervised contents by automatic extraction of facets that are relevant for search result for personal web search as user search interest pattern from text databases. Facet hierarchies are generated for a whole collection, instead of for a given query.

The facets of the query provide an interesting and useful knowledge on a query and, therefore, can be used to improve research experiencesin many ways first; we can show facets of query along with the original search results in an appropriateway. Therefore, users can understand some important aspects ofa query without flipping through dozens of pages. For example, a user can learn different brands and categories of watches.We can also implement a multifaceted search based on facets of mined queries. The user can clarify their specifictry to select facet elements .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.

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

Proposed facets searching system for information discovery and media exploration in online search results. Proposed system extracts and aggregates the useful semantic information from the specific knowledge database Wikipedia. In this paper, A proposed system explore to automatically find query related aspect of search for open-domain queries in Web search engine. 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.

1. important lists. Automatically mining query Facet by clustering from free text and HTML tags in search results. Author further apply fine grained similarity to avoid duplication of list. [10]

I. PREVIOUS WORK

1. Query Reformulation and Recommendation:

Query reformulation and query recommendation (or query suggestion) are two popular ways to help users better describe their information need. Query reformulation is the process of modifying a query that can better match a user's information need and query recommendation techniques generate alternative queries semantically similar to the original query.

2. Query-Based Summarization:

Summarization algorithms are classified into different categories in terms of their summary construction methods (abstractive or extractive), the number of sources for the summary (single document or multiple documents), types of information in the summary (indicative or informative), and the relationship between summary and query (generic or query-based). The difference is that most existing summarization systems dedicate themselves to generating summaries using sentences extracted from documents.

3. Entity Search:

In Existing system, entity Search problem occur. Some existing entity search approaches also exploited knowledge from structure of webpages. Finding query facets differs from entity search in the following aspects.

1. Finding query facets is applicable for all queries, rather than just entity related queries.
2. They tend to return differenttypes of results. The result of an entity search is entities, their attributes, and associated homepages, whereas query facets are comprised of multiple lists of items, which are not necessarily entities.

4. Query Facets Mining and Faceted Search:

Query Faceted search is a way for searching users to find, analyze, and navigate through search data form online web pages. It is widely used in e-commerce and digital libraries. An effective approach for faceted search is the scope of this implementation. Most existing faceted search and facets generation systems are built on a specific domain (such as product search) or predefined facet categories. For example, Web search mining for an unsupervised contents by automatic extraction of facets that are relevant for search result for personal web search as user search interest pattern from text databases. Facet hierarchies are generated for a whole collection, instead of for a given query.

Disadvantages:

1. Extracting query facets using aggregating frequent lists from free text, HTML tags, and repeat regions within top search results.
2. Unique Website model are listed by context similarity in the form of duplicate domain names and published date or republished date.
3. Item ranking for facets are generated by assigning weight to list that contains same facets.
4. Uses semantic similarity model that is addressing the HTML tags for extracting information from webpage.
5. Existing system returns irrelevant data.
6. Existing system contains Cold start problem.

SYSTEM ARCHITECTURE

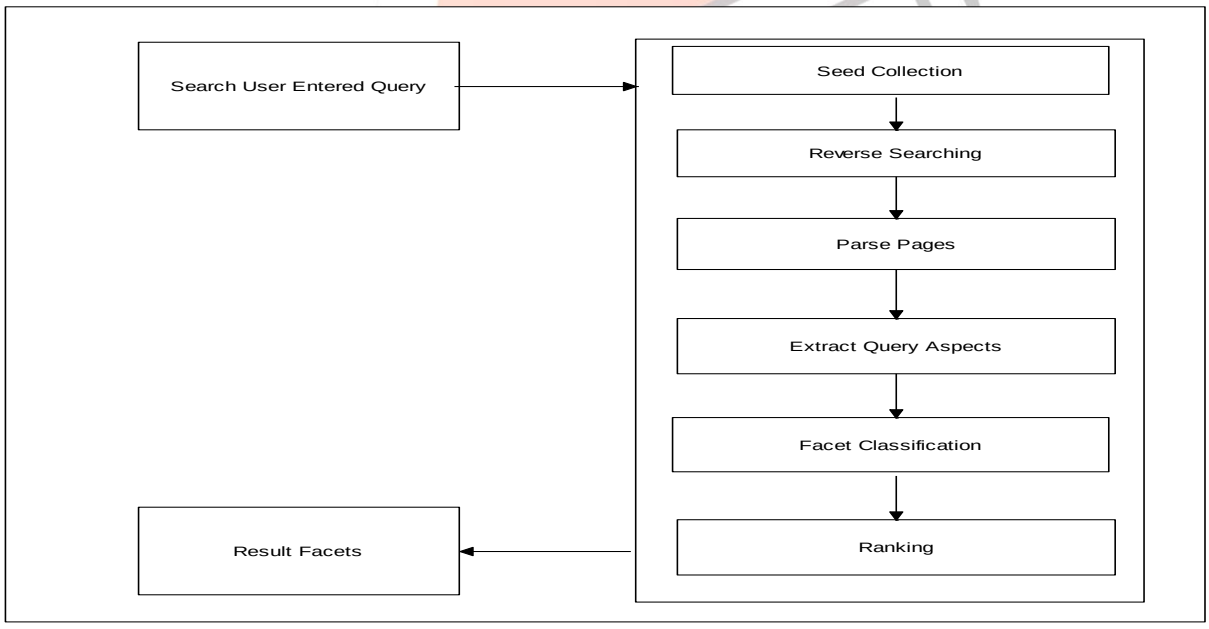

Fig. System Architecture

System Overview:

1. Seed collection:

Here input to system is collect from online API. Which accepts the query and according to query it gives links according to query.

2. Reverse Searching:

Reverse searching is performed to find seeds are relevant to query or not.

5. Unique website identification:

Here unique URL only finds and that unique only passes to next step. We performing this step after getting seeds from seed collection by matching two pagescontent's o for the next step of page parsing will not apply on duplicated links. That will save the time of our system. 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. 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 re-published 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.

6. Parses Pages:

For a list extracted from a HTML element like SELECT, UL, OL, or TABLE by pattern. That contains facet and links that will display to user.

7. **Extract Query aspects from page:**

After performing page extraction we get facets and links. SELECT For the SELECT tag, we simply extract all text from their child tags (OPTION) to create a list. UL/OL For these two tags, we also simply extract text within their child tags (LI). For a list extracted from a HTML element like SELECT, UL, OL, or TABLE by pattern HTMLTAG, its context is comprised of the current element and the previous and next element if any.

8. Facet classification and ranking:

Facets are clustered according to different classes. It cluster data of similar facets and rank the facets good facet should frequently appear in the top results, a facet c is more important.Model (DOM) is applied over html document by parsing html tags. Design fine grained similarity to classify by comparing their similarity. List clustering Similar lists are grouped together to compose a facet. For example different lists about watch gender types are grouped because they share the same items men's and women's.

CONCLUSION

In this paper, we study the problem of finding query facets comparatively faster through suggestion. We propose a systematic solution, which we refer to as QDMiner, to automatically mine query facets by aggregating frequent lists from free text, HTML tags, and repeat regions within top search results. We further analyze the problem of duplicated lists, and find that facets can be improved by modeling fine-grained similarities between lists within a facet by comparing their similarities. To improve performance, we are using log file of generated facets to store it.

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. WebSearch Data Mining, 2008, pp. 33–44.

[2] M. Diao, S. Mukherjea, N. Rajput, and K. Srivastava,, **"Facetedsearch 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 ACMInt. Conf. Inf. Knowl. Manage., pp. 3–12, 2008.

[4] W. Kong and J. Allan, **"Extending faceted search to the generalweb,"** 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 SIGMODInt. Conf. Manage. Data, 2007, pp. 1144–1146.

[6] K. Balog, E. Meij, and M. de Rijke, **"Entity search: Building bridgesbetween 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 forwikipedia,"** in Proc. 19th Int. Conf. World Wide Web, 2010,pp. 651–660.

[9] W. Dakka and P. G. Ipeirotis, **"Automatic extraction of useful facethierarchies from text databases,"** in Proc. IEEE 24th Int. Conf. DataEng., 2008, pp. 466–475.

[10] A. Herdagdelen, M. Ciaramita, D. Mahler, M. Holmqvist, K. Hall,S. Riezler, and E. Alfonseca, "**Generalizedsyntactic and semantic models of query reformulation,"** in Proc. 33rd Int. ACM SIGIRConf. 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 intosnippets for improved query expansion,"** in Proc. 22nd Int. Conf.Comput. Ling., 2008, pp. 737–744.
[14] X. Xue and W. B. Croft, **"Modeling reformulation using query distributions,"**
ACM Trans. Inf. Syst., vol. 31, no. 2, pp. 6:1–6:34, May
2013.
TABLE 9
Results of the Context Similarity Model
fp-NDCG rp-NDCG
UserQ QDMiner 0.631 0.222
Context 0.656 0.248
PageDedup 0.641 0.227
RandQ QDMiner 0.627 0.248
Context 0.664 0.276
PageDedup 0.634 0.252
396 IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 2, FEBRUARY 2016
[15] 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.
[16] J. Huang and E. N. Efthimiadis, **"Analyzing and evaluating query reformulation strategies in web search logs,"** in Proc. 18th ACM Conf. Inf. Knowl. Manage., 2009, pp. 77–86.
[17] 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.
[18] Z. Zhang and O. Nasraoui**, "Mining search engine query logs for query recommendation,"** in Proc. 15th Int. Conf. World Wide Web,2006, pp. 1039–1040.
[19] 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.
[20] I. Szpektor, A. Gionis, and Y. Maarek, **"Improving recommendation for long-tail queries via templates,"** in Proc. 20th Int. Conf. World Wide Web, 2011, pp. 47–56.
[21] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh, **"A comprehensive survey on text summarization systems,"** in Proc. 2nd Int. Conf. Comput. Sci. Appli., 2009, pp. 1–6.
[22] M. Damova and I. Koychev**, "Query-based summarization: A survey,"** in Proc. S3T, 2010, pp. 142–146.
[23] K. Shinzato and T. Kentaro, **"A simple www-based method for semantic word class acquisition,"** in Recent Advances in Natural Language Processing (RANLP '05), pp. 207–216, 2005.
[24] 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.
[25] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, **"Webtables: Exploring the power of tables on the web,"** VLDB,vol. 1, pp. 538–549, Aug. 2008.
[26] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates**, "Web-scaleinformation extraction in knowitall: (preliminary results),"** in Proc. 13th Int. Conf. World Wide Web, 2004, pp. 100–110.
[27] 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.
[28] E. Stoica and M. A. Hearst, **"Automating creation of hierarchical faceted metadata structures,"** in Proc. Human Lang. Technol. Conf., 2007, pp. 244–251.
[29] S. Basu Roy, H. Wang, G. Das, U. Nambiar, and M. Mohania, **"Minimum-effort driven dynamic faceted search in structured databases,"** in Proc. ACM Int. Conf. Inf. Knowl. Manage., 2008, pp. 13–22.
[30] J. Pound, S. Paparizos, and P. Tsaparas, **"Facet discovery for structured web search: A query-log mining approach,"** in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2011, pp. 169–180.