Journal of Engineering and Applied Sciences 13 (3): 575-579, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

Facets Mining for Queries from Search Results

Rasika Langde and Ashwini Zadgaonkar
Department of Computer Science and Engineering,
Shri. Ramdeobaba College of Engineering and Management, 440013 Nagpur, India

Abstract: In this day and age for any query, the user use search engine for their query but every time, relevant information never was accessible on the internet. For a query, there exist vast data and many links available for the user. Time and again, it's been a jigsaw puzzle for most of the users that how to identify and select web links to derive the requisite information. To simplify this, our approach facilitates relevant information by mining the facets that extracting some HTML tags of the content page of Wikipedia. The facets are the keywords or semantically related words which define the query. It displays relevant information or results along with original search results instantly which enhances the user experience of search engine. Wikipedia is trustworthy web portal for the users which gives away results for most of the phenomenon. In this study, we discuss introduction, some preliminary research, our proposed approach, conclusion and future research.

Key words: Facets, facet mining, query, query facet, Wikipedia, information

INTRODUCTION

The facet mining is the part of information retrieval in which information is summarized. The Information Retrieval (IR) is the activity that obtains relevant information from the collection of resources. Search Engine (SE) provides this information via. the internet. Now a days, Google is widely used amongst many of search engines. Currently, SE becomes the vital tools for web users to locate the information. It retrieves relevant digital information from user-fed keywords. Many SEs also provide advanced refining tools (i.e., filters) for users to derive relevant information. Filters differentiate the items and exclude those which don't meet certain criteria in a given set of content. Facets expand the idea of filters. SE gives output in the form of links; some are relevantand some are irrelevant to the query. The user's time is wasted to pop up every linkand hence, the facets play important role in this case. In this study, our approach is to adopt the new system by which user can understand that what type of information is available in this link and web page.

The facets are the keywords in the web pages or set of hyponym relation terms which define the aspect of a query. For a single query, there are multiple or n-number of facets that summate the information of query from the different perspective. A query facet classify the terms by grouping sibling terms together. For example, for the query, Mars landing, there are three possible query facets are shown in below.

Query; Mars landing:

- Curiosity, opportunity, spirit
- USA, UK, Soviet Union
- Video, pictures, news

The first query facet includes {curiosity, opportunity, spirit} which shows different Mars rovers. The second query facet includes {USA, UK, Soviet Union} which displays countries relevant to Mars landings. In these both facets, the terms are instances of the same semantic class while the last one is different than these. The third facet includes {video, pictures, news} which shows the labels for distinct query subtopics.

A query facet can be extracted from search engine results togive asummary about a query that are useful to users and help them to explore the information. The query facets are the terms that share semantic relationship; it gives interesting knowledge about a query and hence can be used to improve search experience in many ways. The query facets can be displayed with original search results. Thus, the users get direct information and instant answer of query. The query facet can be used for query recommendation and query reformulation.

Literature review

Automatic facet mining: The facet mining is important for user to get instant information from search results. Wei *et al*. (2013) and Dou *et al*. (2016) developed QD-miner system to automatically mine query facets by extracting a list from free text, HTML tags and repeat

Corresponding Author: Rasika Langde, Department of Computer Science and Engineering,
Shri. Ramdeobaba College of Engineering and Management, 440013 Nagpur, India

regions contained in the top results and grouped them into a cluster by modified quality threshold algorithm and then calculate weight and rank the facets and facet items, respectively. In this study, they proposed two models (i.e., Unique website model and context similarity model) to rank the query. In unique website model, they recover the problem of the same website which has duplicated information and have identical facets. While in context similarity model, the fine-grained similarity between each pair of lists is recovered. The query dimension and the query facets provide interesting and useful knowledge about query (Wei *et al*., 2013; Dou *et al*., 2016). In "Survey on Query Facets Mining Approaches", they first reviewed the research topics related to representative facets mining task, existing automatic facet extraction methodsand they discussed about the evaluation metrics used for measuring quality and ranking of query facets.

Faceted hierarchies and taxonomies: Faceted search is a technique for accessing information organized according to a faceted classification system, allowing users to digest, analyse and navigate through multidimensional data.

Stoica *et al*. (2007) described the Castanet algorithm to select facet terms based on term frequency distribution. The main idea behind the Castanet algorithm is to get hold of the structure from hypernym relation within the WordNet lexical database (Stoica *et al*., 2007).

Dakka and Ipeirotis (2008) proposed an unsupervised automatic facet term extraction algorithm which is useful for browsing text databases for facets. Further, they work on the expansion of each phrase with "Context" phrases using external resources such as WordNet and Wikipedia. Then compare the term distributions in original database and expanded database to identify terms for browsing facets (Dakka and Ipeirotis, 2008).

Zheng *et al*. (2013) developed a novel system that they called DFT-extractor and DF-miner, respectively in which automatically construct domain-specific faceted taxonomies from Wikipedia in three steps. It first crawls the term of domains from Wikipedia. Further, it exploits a hyponym relation based on the local connectivity pattern of a Wikipedia article graph in which nodes and edges represent the article pages and hyperlinks. It finally constructs a faceted taxonomy by applying a community detection algorithm and a group of a heuristic rules. DFT-extractor also provides a graphical user interface for visualizing the learned hyponym relations and the tree structure of taxonomies (Wei *et al*., 2013, 2015).

In DF-miner system (Wei *et al*., 2015), they proposed a new topology-based approach to automatically acquire domain-specific facets. For this, they worked on Wikipedia article pages and category pages as well as focused on automatic DF-mining based topological properties of hyperlink structure between Wikipedia pages. There is a community structure for the Wikipedia for detection of community. They used the Louvain community detection algorithm by which the different communities have been detected for the graph of Wikipedia.

Mei *et al*. (2007) propose a probabilistic approach which automatically labels multinomial topic models in an objective way. The methods used for labelling are effective to generate the meaningful labels which are useful for interpreting the discovered all topic models such as PLSA, LDA and their variations. Their proposed methods are evaluated using two text data sets with different genres (Mei *et al*., 2007).

In web page clustering based on novel latent semantic approach (Manikaran and Duraiswamy, 2013), they study the alternatives of different types of web clustering information like K-means clustering, spectral clustering. These are used for traditional web page clustering by web page tagging. In information retrieval main trends is how tagged can be used to improve web document clustering and how it is better for user. There are similar types of data in many web documents but by using ranked retrieval it will easier to user to get to what he want. Tagging is beneficial for improving the performance of clustering but the web pages are tagged is restrictive assumption. So, they proposed a new web page grouping approach based on Probabilistic Latent Semantic Analysis (PLSA) mode by which searching time will be reduced. It is based on the likelihood principle and defined proper generative model of data.

MATERIALS AND MEHTODS

Faceted hierarchies and taxonomies: Roy *et al*. (2009) developed a system named DynaCet-a domain-independent middleware system that sites between a user and database provides effective minimum effort based dynamic faceted search solution over enterprise user. At every step, a user is asked one more question about the different facet terms and most promising set of facet terms are identified faceted search based upon the user response (Roy *et al*., 2009).

Kong and Allan (2013) developed a supervised method based on graphical mode for facet query extraction. In this, they proposed two algorithms for approximate interference on the graphical model. The graphical model shows how likely the term should be selected and how likely the two terms should be grouped together (Kong and Allan, 2013).

Pound *et al*. (2011) worked on structured data sources and the existing fundamental problem in supporting faceted search have to be found out an ordered selection of attributes and values that has to be generated the facets. They research on existent structure data and model, the user faceted search by using the intersection of web query logs. They formulate the problem selection of query-log based facet attribute and value (Pound *et al*., 2011).

Ben-Yitzhak *et al*. (2008) addresses the shortcoming of extensions of the faceted search application and enablingusers to gain insight into their data which are far richer than knowing qualities of facets belonging to the document. Then there another extension shows how one can efficiently extend a faceted search engine to support computed facets (Ben-Yitzhak *et al*., 2008).

RESULTS AND DISCUSSION

Proposed approach: In this study, we propose an approach which is based on unsupervised technique that automatically mines the query facets by extracting paragraph and hyperlink tags of HTML. For this approach, we have to be work on Wikipedia web page. The following processhas done on Wikipedia page for facet mining and the block diagram of our approach is given in Fig. 1.

The users write down their query in search engine space andget the search results as a link of Wikipedia which crawls the Wikipedia page. There are two page links of Wikipedia for a query, first is the Wikipedia article page and second is the Wikipedia category page.

From the source page of Wikipedia article and category page, the facet extracts the Paragraph and hyperlink tags of HTML. Much of information is available in paragraph tag and by the link of hyperlink you can get maximum data because hyperlink connects to more pages (Fig.2).

Apply pre-processing on the text page, i.e., the Wikipedia article page from which the paragraph tags extracted. In pre-processing, the stop words and special symbols should be removed and convert the uppercase letters to lowercase letter.

Calculate the TF (Term Frequency) for every term which is obtained after applying pre-processing on Wikipedia article page and Wikipedia category page. For calculating TF, each word should be checked (Fig. 3).

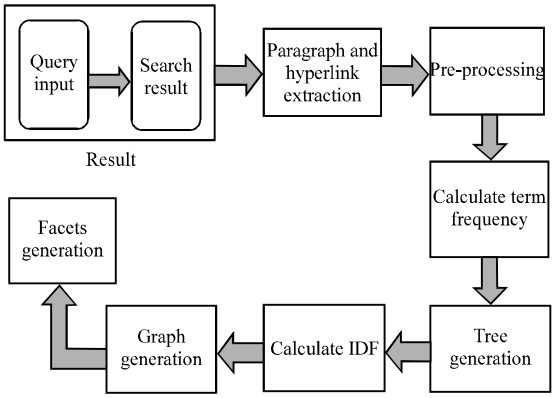
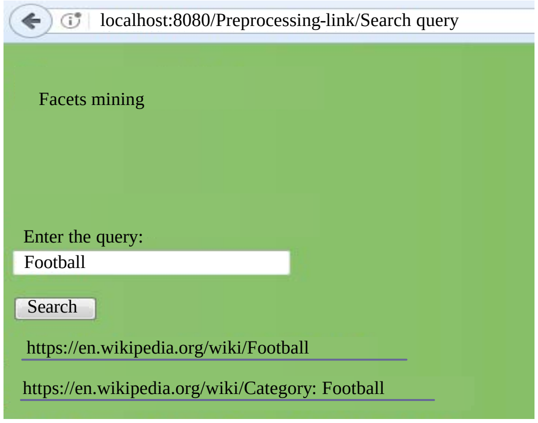


Fig. 1: Block diagram of propose approach



localhost:8080/Preprocessing-link/Search query

Facets mining

Enter the query:

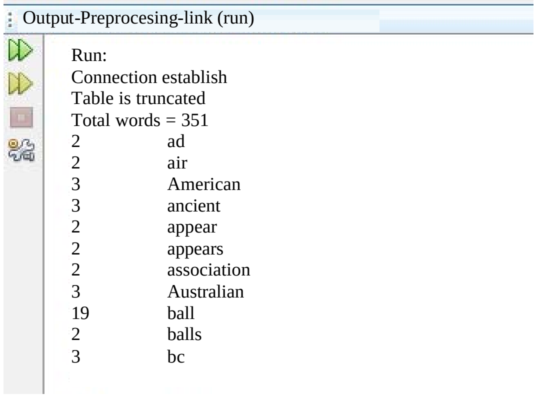
Football

Search

https://en.wikipedia.org/wiki/Football

https://en.wikipedia.org/wiki/Category: Football

Fig. 2: Search query of Wikipedia from search engine



Output-Preprocesing-link (run)

```
Run:
Connection establish
Table is truncated
Total words = 351
2           ad
2           air
3           American
3           ancient
2           appear
2           appears
2           association
3           Australian
19          ball
2           balls
3           bc
```

Fig. 3: Calculated TF

From the calculated TF, the graph will be generated in the form of first level tree structure for Wikipedia article page and Wikipedia category page called as WAT (Wikipedia article tree) and WCT (Wikipedia category tree), respectively (Fig. 4).

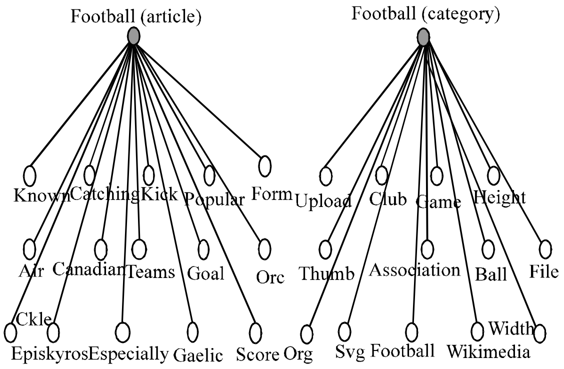


Fig. 4: Calculated TF graph

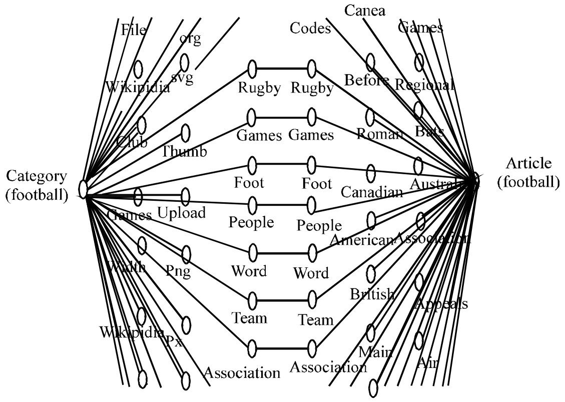
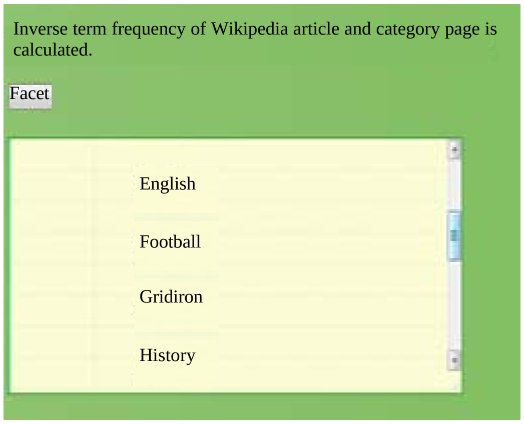


Fig. 5: Calculated IDF graph


Inverse term frequency of Wikipedia article and category page is calculated.

English

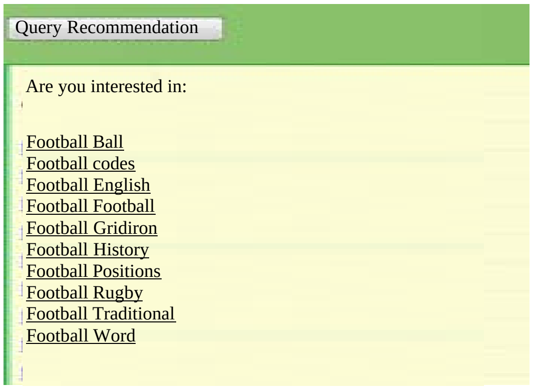
Football

Gridiron

History

Fig. 6: Generated facets

Calculate the common terms from the above generated trees of term frequency and generate the below graph in which the common terms shows from WAT and WCT. Then facets will be generated for Wikipedia page.The generated facets will be used for query recommendation which is helpful to user for better experience (Fig. 5-7).



Are you interested in:

Football Ball
Football codes
Football English
Football Football
Football Gridiron
Football History
Football Positions
Football Rugby
Football Traditional
Football Word

Fig. 7: Query recommendation

CONCLUSION

In this study, we study many papers and the literature review of facet mining. Here, we adopt a new system for facet mining by which users get instant answers about their queries and for this, we worked on Wikipedia's article page and category page. Further, we will research on multiple search result links.

REFERENCES

Ben-Yitzhak, O., N. Golbandi, N. Har'El, R. Lempel and A. Neumann *et al*., 2008. Beyond basic faceted search. Proceedings of the 2008 International Conference on Web Search and Data Mining, February 11-12, 2008, ACM, Palo Alto, California, ISBN:978-1-59593-927-2, pp: 33-44.

Dakka, W. and P.G. Ipeirotis, 2008. Automatic extraction of useful facet hierarchies from text databases. Proceedings of the IEEE 24th International Conference on Data Engineering (ICDE'08), April 7-12, 2008, IEEE, Cancun, Mexico, ISBN:978-1-4244-1836-7, pp: 466-475.

Dou, Z., Z. Jiang, S. Hu, J.R. Wen and R. Song, 2016. Automatically mining facets for queries from their search results. IEEE. Trans. Knowl. Data Eng., 28: 385-397.

Kong, W. and J. Allan, 2013. Extracting query facets from search results. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28-August 01, 2013, ACM, Dublin, Ireland, ISBN:978-1-4503-2034-4, pp: 93-102.

Manikaran, P. and K. Duraiswamy, 2013. Web page clustering based on novel latent semantic approach. Intl. J. Soft Comput., 8: 149-153.

Mei, Q., X. Shen and C. Zhai, 2007. Automatic labeling of multinomial topic models. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 12-15, 2007, ACM, San Jose, California, ISBN: 978-1-59593-609-7, pp: 490-499.

Pound, J., S. Paparizos and P. Tsaparas, 2011. Facet discovery for structured web search: A query-log mining approach. Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data, June 12-16, 2011, ACM, Athens, Greece, ISBN:978-1-4503-0661-4, pp: 169-180.

Roy, S.B., H. Wang, U. Nambiar, G. Das and M. Mohania, 2009. Dynacet: Building dynamic faceted search systems over databases. Proceedings of the 2009 IEEE 25th International Conference on Data Engineering, March 29-April 2, 2009, IEEE, Arlington, Texas, ISBN:978-1-4244-3422-0, pp: 1463-1466.

Stoica, E., M.A. Hearst and M. Richardson, 2007. Automating creation of hierarchical faceted metadata structures. Proceedings of the Human Language Technologies 2007: North American Association for Computational Linguistics (HLT-NAACL'07), April 22-27, 2007, Association for Computational Linguistics, Rochester, New York, pp: 244-251.

Wei, B., J. Liu, J. Ma, Q. Zheng and W. Zhang *et al*., 2013. DFT-extractor: A system to extract domain-specific faceted taxonomies from Wikipedia. Proceedings of the 22nd International Conference on World Wide Web, May 13-17, 2013, ACM, Rio de Janeiro, Brazil, ISBN: 978-1-4503-2038-2, pp: 277-280.

Wei, B., J. Liu, Q. Zheng, W. Zhang and C. Wang *et al*., 2015. Df-miner: Domain-specific facet mining by leveraging the hyperlink structure of Wikipedia. Knowl. Based Syst., 77: 80-91.

Zheng, B., W. Zhang and X.F.B. Feng, 2013. A survey of faceted search. J. Web Eng., 12: 041-064.