International Journal of Engineering & Technology, 7 (1) (2018) 121-125

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i1.8944
Research paper

Facets extraction-based approach for query recommendation using data mining approach

A. V. Zadgaonkar [1]*, A. J. Agrawal [2], S. Aote [3]

[1] *Assistant Professor, CSE Dept, Ramdeobaba College Of Engineering & Management, Nagpur, India - 440013*
[2] *Associate Professor CSE Dept, Ramdeobaba College Of Engineering & Management, Nagpur, India – 440013*
[3] *Assistant Professor, CSE Dept, Ramdeobaba College Of Engineering & Management , Nagpur, India - 440013*

**zashwini@rediffmail.com:*

Abstract

Search engines are popularly utilized for extracting desired information from World Wide Web by users. Efficiency of these search engines are dependent on how fast search results can be retrieved and whether these results reflects the desired info or not. For a particular query, vast amount of relevant information is scattered across the multiple web pages. Search engines generate multiple web links as a output. It has been a jigsaw puzzle for users to identify and select relevant links to extract further desired information. To address this issue, we are proposing an approach for Query Recommendation for getting relevant search results from web using facet mining techniques. Facets are the semantically related words for a query which defines its multiple aspects. We are extracting these aspects of a query from Wikipedia pages which is considered to be a trustworthy resource on the web. Our proposed system uses various text processing techniques to refine the results using lexical resource like WorldNet. In this paper we are discussing our approach and its implementation and results obtained. In the paper , Discussion on future research direction is included to conclude.

Keywords: *Facets, Information Retrieval, Query Recommendation*

1. Introduction

The information retrieval (IR) activity obtains relevant information from the collection of resources about a query. Facet mining is used for information retrieval task where results are produced in categorized form which can be further selected by user for intended information. Search Engines like Google plays the vital role for web users to locate the desired information. It retrieves relevant information from user-feed keywords. Many Search Engines provide advanced refining tools (i.e. Filters) for users to derive relevant information. Filters differentiate and exclude those which does not meet certain criteria's in a given set of contents. Facets expand the idea of filters. Search Engines generates gives output in the form of web links. Some of them may be relevant, and some may be irreverent. Here user time is wasted in exploring each and every link. Facets can play an important role as a filter. Our proposed approach is about recommending users about query by giving suggestions for further exploration.

The facets are the keywords or set of hyponym relation terms which defines different aspects of a query. For a single query, there can be n-number of facets which summarizes query information from the different perspectives/aspects. A query facet classifies the terms by grouping sibling terms together. For example, the query, ***Mars landing***, has three possible facets as

Query: ***mars landing***
1. *Curiosity, Opportunity, Spirit*
2. USA, UK, Soviet Union
3. Video, pictures, news

The first facet group includes {Curiosity, Opportunity, Spirit} which relates with different Mars rovers. The second facet group includes {USA, UK, Soviet Union} which are the country names relevant to Mars landings. In these both groups, the terms are instances of the same semantic class while the last group is different. The third facet group includes {video, pictures, news} which are the labels for distinct query subtopics.

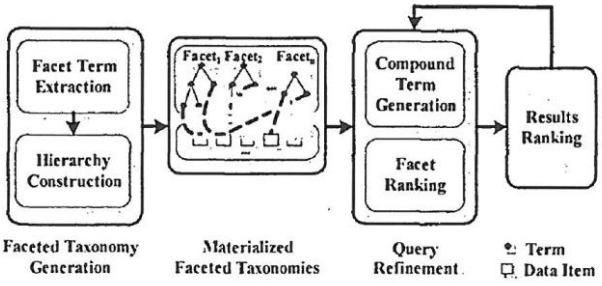


Fig 1: General framework for faceted search system

Query facets can be extracted from search engine results to give summary about a query. It is useful to the users for exploring the desired information. The query facets are the phrases or terms that share semantic relationship; it gives interesting knowledge about query and hence can improvise search experience in many ways. The query facets can be generated along with

Copyright © 2016 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

original search results. The query facet can be used further for query recommendation and query reformulation tasks.

2. Related Work

Keyword base searching techniques are popularly used among internet users for extracting desired info from net. Following table shows the comparative analysis of keyword based techniques with faceted search on various parameters

Items	Faceted Search	Keyword Search
Search Interface	**Faceted taxonomy:** 1. Using multidimensional taxonomies for satisfying variant search needs. 2. Dynamic taxonomy 3. Using a mouse clicking for navigations.	**Keyword:** Suffering from "vocabulary problem"
Prior Knowledge	Requiring little prior knowledge of data schema	Requiring prior knowledge of the dataset to be searched
Navigation Function	1. Refining search results using different facets 2. The number of data items in each category can be used for next navigation.	"Trial-and-error" interaction having no navigation function
Diversification Function	Diversifying search results using only a small number of facet terms	"One result list only", not supporting search results diversification
Ranking Function	Supporting facet ranking and search results ranking	Only supporting search ranking

Table 1:comparison between faceted search and keyword based search

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. Automatic Facets Mining is the exploratory search mechanism which provides an iterative way to refine search results by a faceted search and faceted taxonomy. It is important for user to get instant information from search results .Dau, Jiang et.al.[1,2] developed *QDMiner* system to automatically mine query facets by extracting a list from free text, HTML tags and repeat 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 the paper, 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.

Stoica [4], described the Castanet algorithm to select facet terms based on term frequency distribution. The main idea behind Castanet algorithm is to get hold of the structure from hypernym relation within the WordNet lexical database.

Dakka and Ipeirotic [5], 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 they compare the term distributions in original database and expanded database to identify terms for browsing facets.

Wei and Liu [6, 7], developed a novel system called DFT-Extractor and DF-Miner respectively which that they 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.

In DF-Miner system [7], they proposed a new topology 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, Xhen and Zhai [8], 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.

Roy and Wang [9], 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.

Kong and Allan [10], 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.

Pound, Paparizo and Tsaparas [11], worked on structured data sources and the existing fundamental problem in supporting faceted search that have to find out an ordered selection of attributes and values that to generate facets. They work on existent structure data and model, and hence 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.

Yitzhah [12], addresses the shortcoming of extensions of the faceted search application and enabling users 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.

3. Methodology

Our aim is to Design a solution for automatically mining query facets by extracting and grouping search results dynamically. To address this problem We are proposing an unsupervised approach for automatically mining query facets by extracting paragraph and hyperlink tags of web pages extracted by search engine. Our system is designed considering Wikipedia as a main resource. Our proposed system architecture can be visualized by the following block diagram.

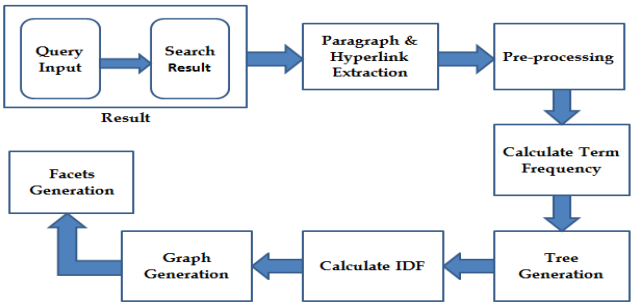


Fig1: Block diagram of Proposed Approach

The system works in the following phases:

3.1 Query Input

- The system collects search results generated by Search Engines. Out of n no of web link generated, our system focuses on Wikipedia link as Wikipedia is considered to be trustworthy resource available on web. Another reason for the preference is time efficiency.

- The Wikipedia link is taken as input for next stage. Our system extracts the paragraph and hyperlink tag from Wikipedia page The paragraph tag extracted from *Wikipedia's Article page* and the hyperlink tag is extracted from *Wikipedia's Category page*.

3.2 Query Preprocessing

- The unstructured information contain lots of noisy data i.e. stop words, special symbols. Such words need to be filtered out before further processing .. In the next step system apply pre-processing on the extracted info for removal of special symbols and stop words.

3.3 Calculation of Term Frequency

- After preprocessing the words are written in words.txt and cat.txt file. These words are used to calculate their term frequency (TF) i.e. how many times the single word occur in the document or file which will be stored in the database for further computation.

3.4 Generation Of Article tree and category tree

- In the next step system generates Wikipedia Article Tree and Wikipedia Category Tree from the terms stored in the database.
- To remove the redundancy in both trees generated common terms are filtered.

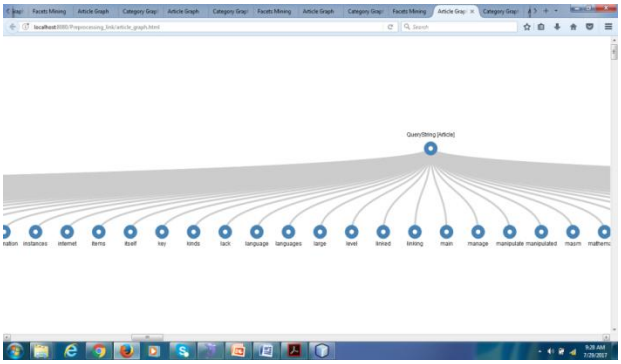


Fig 2: Wikipedia Article tree

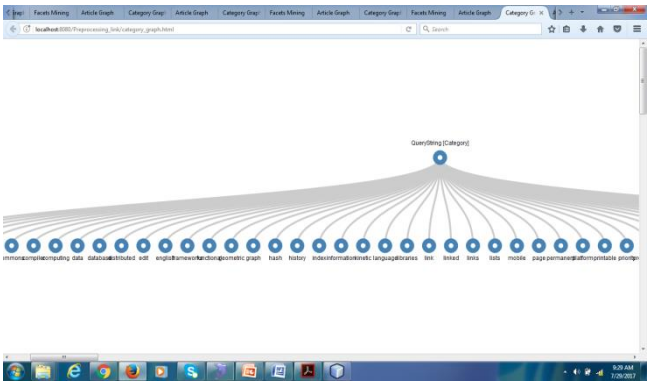


Fig 3: Wikipedia Category tree

3.5 Facet Generation

- To generate facets for a given search queries our system extracts synonyms and hyponyms of the terms. We are using WorldNet dictionary which a commonly used Lexical resource. As a final output our system generates facets for a given query that can be for query recommendations.

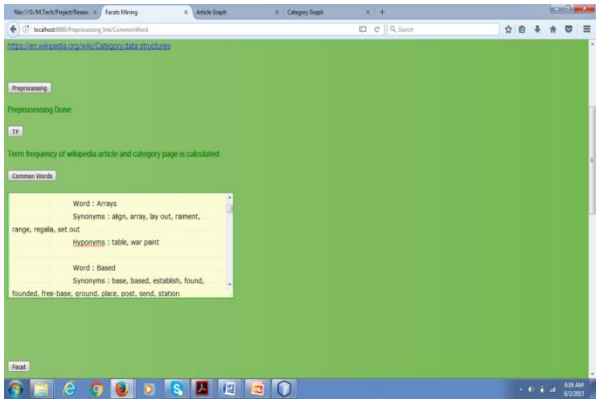


Fig 4: Proposed system interface

4. Resources and Tools Required

- **WordNet**

WordNet (Fellbaum 1998) [5] is a semantic lexicon for the English language developed and maintained at the cognitive science laboratory of Princeton University, New Jersey and it as developed under the direction of Professor George Miller. There are 206941 words across 117659 SynSets, which are groups of synonyms, in WordNet 3.0. This system uses wordnet for getting the synonyms and hyponyms of the words, which are facets.

- **JAWS Library**

Java API for WordNet searching (JAWS) is a library that we use for retrieving relations from wordnet lexical database. It is maintained by CSE department at Southern Methodist University. It is used for providing knowledge of ambiguous terms by retrieving information associated with it i.e. its word form, phrases etc.

- **Lucene**

Apache Lucene is a free and open-source information retrieval software library, originally written in Java by Doug Cutting. It is supported by the Apache Software Foundation and is

released under the Apache Software License. Lucene has been ported to other programming languages including Delphi, Perl, C#, C++, Python, Ruby, and PHP.

5. Results and Discussion

Our system focuses on two Wikipedia links https://en.wikipedia.org/wiki/andhttps://en.wikipedia.org/wiki/Category for facet extraction purpose. These pages are processed by clicking on these links to extract the paragraph tag …. from the Wikipedia Article page and extract the hyperlink tag<a>…. from Wikipedia Category page. These web pages contents are in unstructured format. For stop words removal, we used Stanford Dictionary stop words list which consists of 470 stop words. The outputs of preprocessing include the removal of stop words. After preprocessing the words are written in words.txt and cat.txt file. These words are used to calculate their term frequency (TF) i.e. how many times the single word occur in the document or file.

Term Frequency, measures how frequently a term occurs in a document. Since every document is different in length, it is possible that a term would appear much more times in long documents than shorter ones. The term frequency (TF) can be calculated as the number of terms t appear in document divided by the total number of terms in that document

$$TF = \frac{Number\ of\ times\ t\ appear\ in\ document}{Total\ number\ of\ terms\ in\ that\ document}$$

After calculating TF, the values of those tf will be store in the database for further use. The database values are used for generating the Wikipedia Article Tree (WAT) and Wikipedia Category Tree (WCT). Two types of the hyperlink structure of Wikipedia are employed to discover Facets. The hyperlink structure of article pages is called Wikipedia Article Tree (WAT). The domain-specific WAT is composed of the article pages related to a domain and the hyperlinks among them. Each article page has a page title that can be considered as a domain-specific term. The hyperlink structure of the category pages is a tree like structure , called Wikipedia Category Tree (WCT). Each category page has a hyperlink pointing to an article page of the same name. For example, the category Data structures corresponds to the article Data structure. From these two generation of trees the common words from those will be extracted and calculated.

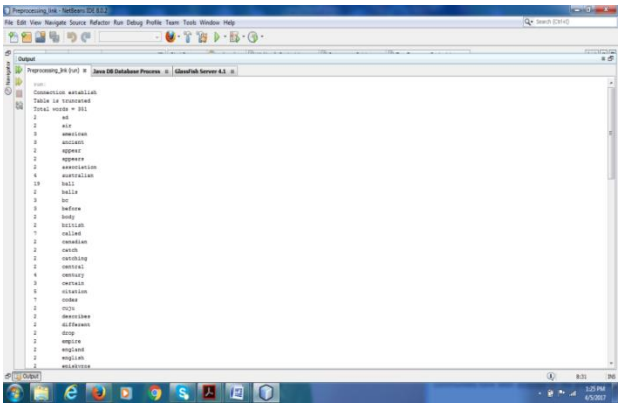


Fig 5: Calculated term frequency

The Facets are generated by the combination of synonyms and hyponyms, which shows the aspects of user query. From the common words, extract the synonyms and hyponyms from the WordNet Dictionary. The facets are generated by the combination of the synonym and hyponyms.

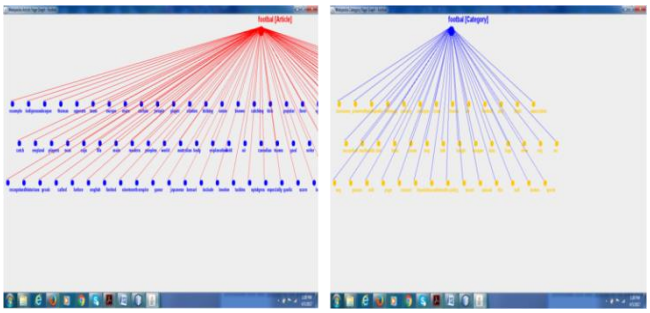


Fig 6: Calculated TF Graph

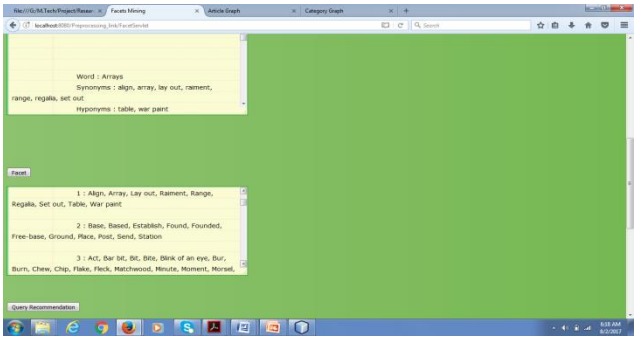


Fig 7: Generated facets

In this Paper, We have presented a new system for facet mining by which users get instant answers about their queries. For this, we worked on Wikipedia's article page and category pages. Following table shows the F-score value calculated for various domain specific terms.

Table 1: Performance Evaluation

Sr. No.	Query	F- Score Average
1.	Computer Science: { Data transmission, Data structures , Natural language processing }	67.03%
2.	Sports: {Football, Cricket, Badminton}	52.52%
3.	Courses: {Engineering, Medical, Filmmaking}	49.24%

The above table shows the analysis of Facets Mining in which average is calculated for some specific domains like queries are executed here for computer science domain. The average is calculated by finding the precision, recall and F-measure of that query domain.

6. Conclusion & Future Scope

To contribute to the domain of IR, we have developed and implemented an easy-to-use system for facet mining by which user get instant answer about their queries covering its all aspects. This system is capable of automatically extracting the facets from Wikipedia resource through Wikipedia Article page and Wikipedia Category page. By calculating TF, the WAT and WCT is generated. Further by removing redundancy and integrating our system with a lexical resource WordNet, our system generates response to user queries in the form of facets.

We want to further explore our work to improve the search experience by providing better query recommendations. Some of the identified future research directions could be as follow.

- To test the system for variety of trusted web resources other than Wikipedia
- To overcome the redundancy issues encountered for some queries of documents and facets.
- To test the system performance on variety of large data sets

References

[1] B. Wei, J. Liu, Q. Zheng, W. Zhang, X. Fu, B. Feng, A survey of faceted search, J. Web Eng. 12 (1–2) (2013) 041–064.

[2] Zhicheng Dou, Member, IEEE, Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song, Automatically Mining Facets for Queries from Their Search Results," IEEE Transactions on knowledge and data engineering, pp. 385-397, 2016.

[3] Sheetal Sonwane1, Nilam Patil2 Survey on Query Facets Mining Approaches, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 , Volume 6 Issue 1, January 2017. www.ijsr.net.

[4] E. Stoica, M.A. Hearst, M. Richardson, Automating creation of hierarchical faceted metadata structures, in: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Rochester, NY, 2007, pp. 244–251.

[5] W. Dakka, P.G. Ipeirotis, Automatic extraction of useful facet hierarchies from text databases, in: 2008 IEEE 24th International Conference on Data Engineering, Cancun, Mexico, 2008, pp. 466–475.

[6] B. Wei, J. Liu, J. Ma, Q. Zheng, W. Zhang, B. Feng, DFT-extractor: a system to extract domain-specific faceted taxonomies from Wikipedia, in: Proceedings of the 22nd International Conference on World Wide Web companion, Rio de Janeiro, Brazil, 2013, pp. 277–280.

[7] Bifan Wei, Jun Liu , Qinghua Zheng , Wei Zhang , Chenchen Wang, Bei Wu, DF-Miner: Domain-specific facet mining by leveraging the hyperlink structure of Wikipedia, Available online 13 January 2015, Knowledge-Based Systems 77 (2015) 80–91.

[8] Q.Mei, X. Shen, C. Zhai, Automatic labeling of multinomial topic models, in: 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, 2007, pp. 490–499.

[9] S.B. Roy, H.D. Wang, U. Nambiar, G. Das, M. Mohania, DynaCet: building dynamic faceted search systems over databases, in: 2009 IEEE 25th International Conference on Data Engineering, Shanghai, China, 2009, pp. 1463–1466.

[10] 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

[11] J. Pound, S. Paparizos, P. Tsaparas, Facet discovery for structured web search: a query-log mining approach, in: 2011 ACM SIGMOD International Conference on Management of Data, Athens, Greece, 2011, pp. 169–180.

[12] 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.