ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017

Automatically Mining Query Facet for Online Content Mining

Akriti Verma[1], Bhagyashri Bhamare[1], Dipak Sanap[1], Ravindra Gorde [1], Prof.Tejaswini Adikane[2]

Student, Department of Computer Engineering, Sinhgad Institute of technology, Lonavala, Savitribai Phule Pune University, Pune, India[1]

Professor, Department of Computer Engineering , Sinhgad Institute of technology, Lonavala, Savitribai Phule Pune University, Pune, India[2]

ABSTRACT: Automatically query facet generation in online searching or content retrieval incritical task. Query facet helps to search movie, video, or shopping from different shopping portal. Proposed work use QD Miner for extracting facet for searchresult from user interest mining. Online search recommendation needs user perceptionabout items to be mined. This work emphasizes facet mining by document content extraction. QD miner classify user search with group analysis from content retrieved. Query mining dig to generate review percentage by user review generation by summarizing user comment about item. Department Keywords.

KEYWORDS: Query facet, Faceted search, Summarization, User intent.

I. INTRODUCTION

A query facet is a set of items which describe and summarize one important aspect of a query. Query facets provide interesting and useful knowledge based upon user interest. Searches for social media sites and shopping sites provide the interest of user. Online shopping is new concept for social media.Online shopping site prefer using search result and product review to buy things from internet. To determine user interest for the product, significant work assigned to shopping websites. We analyze that important aspect of information about a query are usually presented in list styles and repeated many times among top retrieved documents. Thus we propose collection for frequent search items within the top search results to mine query facets and implement a system called QD Miner. Facet rank is dependent on the unique website and their lists is not convincing. Query facets contain structured knowledge covered by the query, they can be used in other fields such as semantic search or entity search. Hence we propose the Context Similarity Model, in which we model the filtered similarity between each pair of product. Summarize user review and generate rating to support product for online shopping to user.

A.BACKGROUND

Proposed system designed to implement web mining for searching query facet to mine movie, video, product search online or offline from relevant data. Relevant facets can be searched by using content mining. Content classification leads to efficiently item search from movie or product items. Extracting user rating and generate review is problem for online searching for product with help of user search interest. Mining useful pattern for search recommendation using user view point by mining user interest i.e. predicting user interest and generate review

B.MOTIVATION

The challenges come from the large and heterogeneous nature of the web, which makes it difficult to generate and recommend facet. The query facet contains a group of words and phrases that summarize the information about query. The information of facets subtopic is not clearly explained therefore in this paper we propose this technique effectively. Previous models typically generate words and phrases related to the original query, but do not consider how these words and phrases would fit together in actual queries. A set of reformulated queries is generated by using a passage analysis technique on the target corpus. The general idea of this technique is based on the observation that passages

DOI: 10.15680/IJIRCCE.2017. 0505200

containing all query words or most of the query words provide a good source of information for query segmentation and substitution. QD Miner aims to offer the possibility of finding the main points of multiple documents and thus save users time on reading whole documents.

C.OBJECTIVE AND GOAL

To overcome the problem of duplication in the lists. Many websites contain the same information and that information is re-published by other websites. Duplication of data is presented in all lists. The idea of transforming the original query into a distribution of actual reformulated queries is motivated by the availability of large scale query logs. It is achieved with the sequence of hidden nodes representing the latent topics of the corresponding terms.

Thus we propose aggregating frequent lists within the top search results to extract query facets and implement a system called QD Miner. The QD Miner extracts lists from free text, HTML tags, and repeat regions contained in the top search results, groups them into clusters based on the items they contain, then ranks the clusters and items based on how the lists and items appear in the top results. We propose two models, the Unique Website Model and the Context Similarity Model, to rank query facets.

II.LITERATURE SURVEY

[A]*"Query-Based Summarization: A survey".M. Damova and I. Koychev.2010:*

Query facets are a specific type of summaries that describe the main topic of given text. The difference is that most existing summarization systems dedicate themselves to generating summaries using sentences extracted from documents, while system generate summaries based on frequent lists.

[B]. *"Entity search: building bridges between two worlds". K. Balog, E. Meij, and M. de Rijke, 2016:*

Mining query facets is related to entity search as for some queries, facet items are kinds of entities or attributes. Some existing entity search approaches also exploited knowledge from structure of webpages. 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.

III.SOFTWARE REQUIREMENT SPECIFICATION

Hardware Resources Required	**Software Resources Require**
Processor– Pentium IV	Operating System: Windows 7/8
Speed - 2.4 GHZ	PROGRAMMING LANGUAGE: JAVA
RAM - 3 GB(Min)	DATA BASE: MSSQL
Hard Disk - 80 GB	TOOL: Eclipse

IV.IMPLEMENTATION STATUS

As per the requirement of institute the system implementation is completed in April 2017.According to system development plan the system is in executable and ready for use. It meets the minimum specified requirements and also upgraded graphical user interface is being made to user to make system user friendly. As the system consisting of web application it has been tasted with multiple input facets.

V.COMPARISON BETWEEN EXISTING SYSTEM AND PROPOSED SYSTEM

Existing System	Proposed System
There are some challenges that the existing approaches have to face in finding both relevant and diverse subtopics, such as term mismatch and data sparseness.	A novel semantic representation for query subtopics is introduced, which including phrase embedding representation and query category distributional representation.

VI.ALGORITHM FOR RELEVANT FEATURE DISCOVERY

Algorithm and Technique:

Classification algorithm (QT Algorithm):

Steps:

1. A random gene is chosen from the selected gene list.
2. The algorithm determines which gene has the greatest similarity to this gene. If their totaldiameter does not exceed the threshold diameter, then these two genes are clustered together.
3. Other genes that minimize the increase in cluster diameter are iteratively added to this cluster. This process continues until no gene can be added to this first candidate cluster without surpassing the diameter threshold.
4. A second gene is chosen.
5. The algorithm determines which gene has the greatest similarity to this second gene. All genes in the selected gene list are available for consideration to form the second candidate cluster.
6. Other genes from the selected gene list that minimizes the increase in cluster diameter are iterativelyadded to the second candidate cluster. The process continues until no gene can be added to thissecond candidate cluster without surpassing the diameter threshold.
7. The algorithm iterates through all genes on the selected gene list and forms a candidate cluster withreference to each gene. In other words, there will be as many candidate clusters as there are genes inthe gene list. Once a candidate cluster is formed for each gene, all candidate clusters below the userspecifiedminimum size are removed from consideration.
8. The largest remaining candidate cluster, with the user-specified minimal number of gene member, is selected and retained as a QT cluster. The genes within this cluster are now removed from consideration. All remaining genes will be used for the next round of QT cluster formation.
9. The entire process (step 1 to 9) is repeated until the largest remaining candidate cluster has fewer than the user-specified number of genes.
10. The result is a set of non-overlapping QT clusters that meet quality threshold for both size, with respect to number of genes, and similarity, with respect to maximum allowable diameter.
11. Genes that do not belong in any QT clusters (as well as genes that are not in the selected gene list) will be grouped under the “unclassified” group.

VII.SYSTEM ARCHITECTURE

Internal software data structure

Data structures that are passed among components the software are described. The java. sql package defines an interface called Java. sql. Driver that makes to be implemented by all the JDBC drivers and a class called java.sql.Driver Manager that acts as the interface to the database clients for performing tasks like connecting to external resource managers, and setting log streams.

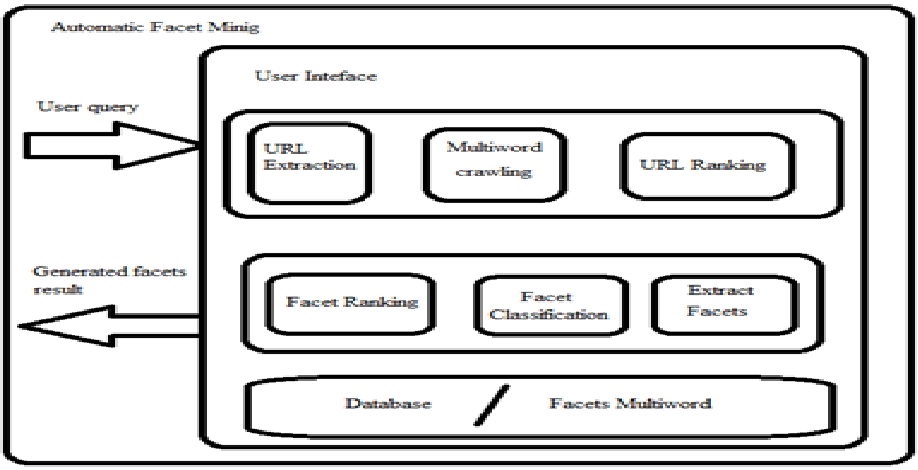


Figure 1: System Architecture

When a JDBC client requests the Driver Manager to make a connection to an external resource manager, it delegates the task to an appropriate driver class implemented by the JDBC driver provided either by the resource manager vendor or a third party

VIII.MATHEMATICAL MODULE

A] Set Theory

Let us consider S as a system for automatically facet mining for shopping portal

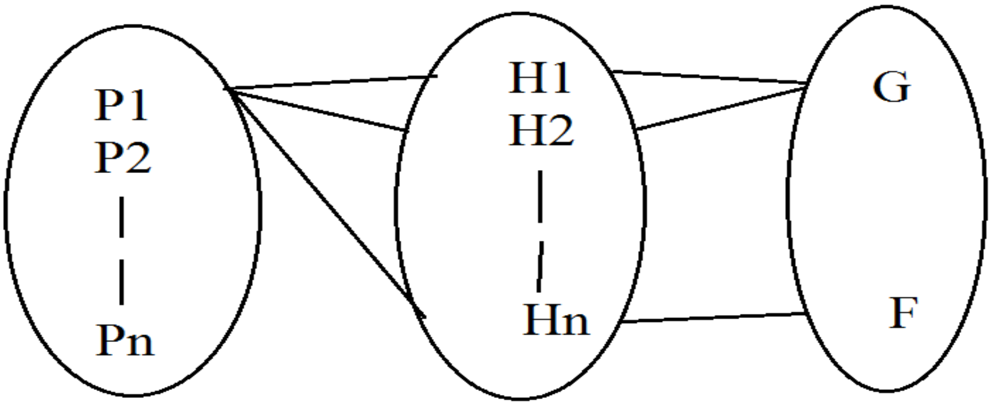
S=

INPUT:

Identify the inputs

F={ f1, f2, f3 ....., fn— F as set of functions to execute commands.}

I={ fi1, i2, i3—I sets of inputs to the function set}



O= {o1, o2, o3.—O Set of outputs from the function sets}

S= {I, F, O}

I = {Query submitted by the user,}

O ={Output of desired query,}

F = {

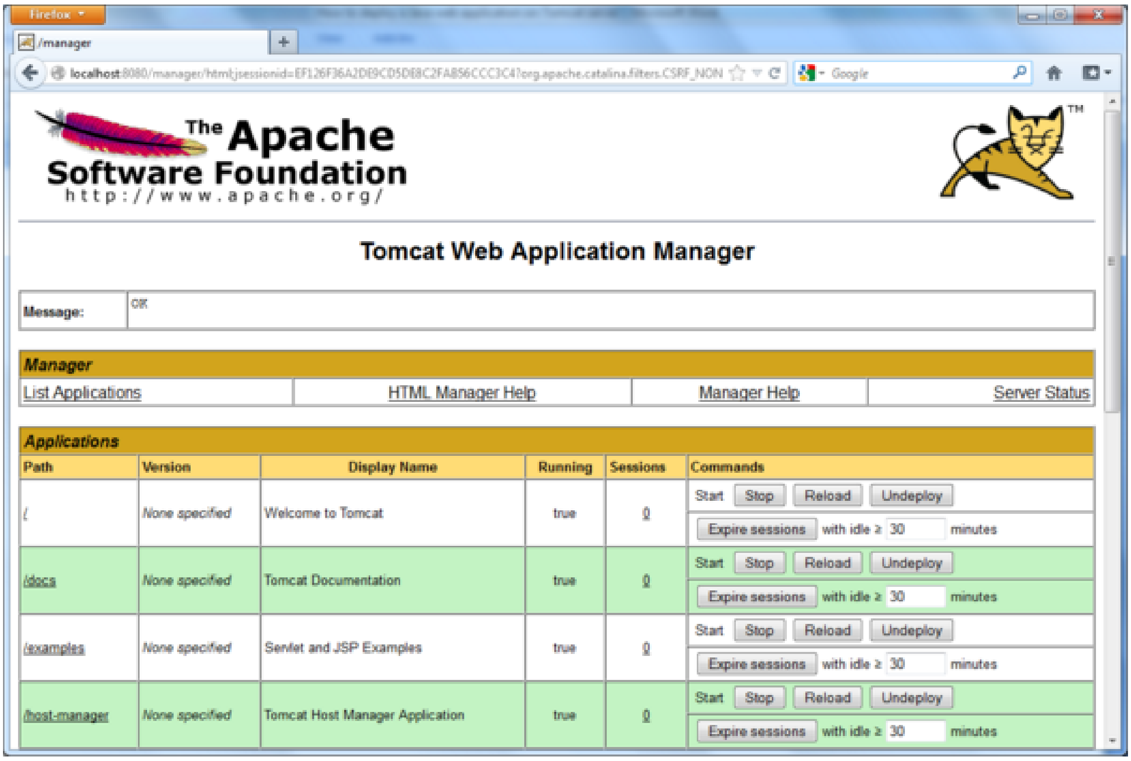
Functions implemented to get the output,

User interest mining,

Above mathematical model is NP-Complete.

IX.EXPERIMENTAL SET UP AND RESULT TABLE

Given below are some of sample images from the existing system that states the experimental results and set up.


Firefox
/manager
localhost:8080/manager/html;jsessionid=EF126F36A2DE9CD5DE8C2FA856CCC3C4?org.apache.catalina.filters.CSRF_NON
Google
The Apache Software Foundation
http://www.apache.org/
TM

Tomcat Web Application Manager

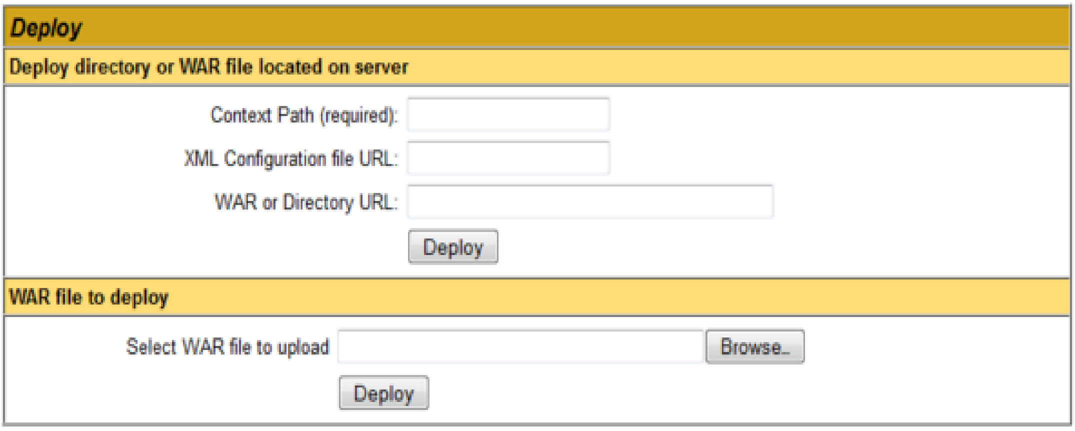
Message: OK

Manager

List Applications | HTML Manager Help | Manager Help | Server Status

Applications

Path	Version	Display Name	Running	Sessions	Commands
/	None specified	Welcome to Tomcat	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/docs	None specified	Tomcat Documentation	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/examples	None specified	Servlet and JSP Examples	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/host-manager	None specified	Tomcat Host Manager Application	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes



Deploy

Deploy directory or WAR file located on server

Context Path (required):

XML Configuration file URL:

WAR or Directory URL:

Deploy

WAR file to deploy

Select WAR file to upload Browse...

Deploy

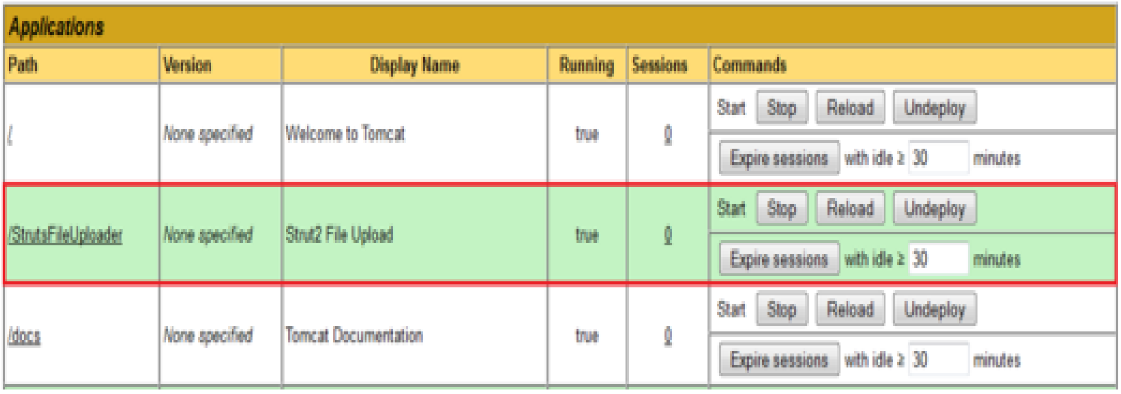
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 5, May 2017



Applications

Path	Version	Display Name	Running	Sessions	Commands
/	None specified	Welcome to Tomcat	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/StrutsFileUploader	None specified	Strut2 File Upload	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes
/docs	None specified	Tomcat Documentation	true	0	Start Stop Reload Undeploy Expire sessions with idle ≥ 30 minutes

X.CONCLUSION

Search item mining through the user search interest using review mining or user rating for product, which can be done by QDMiner, for effectively mine query facetsby searching frequent user review for the product from online searching, HTML tags, and user comment are considered to generate review about product from topsearch. In proposed system combined metrics to evaluate the quality of query facets.Experimental results show that useful query facets are mined by the approach. Wefurther analyze the problem of duplicated lists, and find that facets can be improvedby modelling fine-grained similarities between lists within a facet by comparing theirsimilarities.

REFERENCES

[1.] R. S. Pressman, Software Engineering (3rd Ed.): A Practitioner's Approach. New York, NY, USA: McGraw-Hill, Inc., 1992.

[2.] Extracting Query Facets from Search Results: Weize Kong and James Allan.

[3.] Query Subtopic Mining by Combining Multiple Semantics: Lizhen Liu, Wenbin Xu, Wei Song, HanshiWang and Chao Du.

[4.] Search Result Diversification Based on Query Facets: Sha Hu, Zhi-Cheng Dou, Xiao-Jie Wang.

[5.] O. Ben-Yitzhak, N. Golbandi, N. HarEl, 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. 3344.

[6.] 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. 312, 2008. Department

[7.] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating queryreformulation strategies in web search logs," in Proc. 18th ACMConf. Inf. Knowl. Manage., 2009, pp. 77–86.

[8.] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query recommendationusing query logs in search engines," in Proc. Int. Conf. CurrentTrends Database Technol., 2004, pp. 588–596.

[9.] Z. Zhang and O. Nasraoui, "Mining search engine query logs forquery recommendation," in Proc. 15th Int. Conf. World Wide Web,2006, pp. 1039–1040.

[10.] L. Li, L. Zhong, Z. Yang, and M. Kitsuregawa, "Qubic: An adaptiveapproach to query-based recommendation," J. Intell. Inf. Syst.,vol. 40, no. 3, pp. 555–587, Jun. 2013.

DOI: 10.15680/IJIRCCE.2017. 0505200