International Journal of Computational Intelligence Research
ISSN 0973-1873 Volume 12, Number 2 (2016), pp. 163-167

http://www.ripublication.com

Simple Web Browsing Through Generated Facets

Mrs. Vaishakhi V K[1] and Mrs. Neethu T Regi[2]

[1]Malabar Institute of Technology, Anjarakandy.
E-mail: vaishakhivk@gmail.com

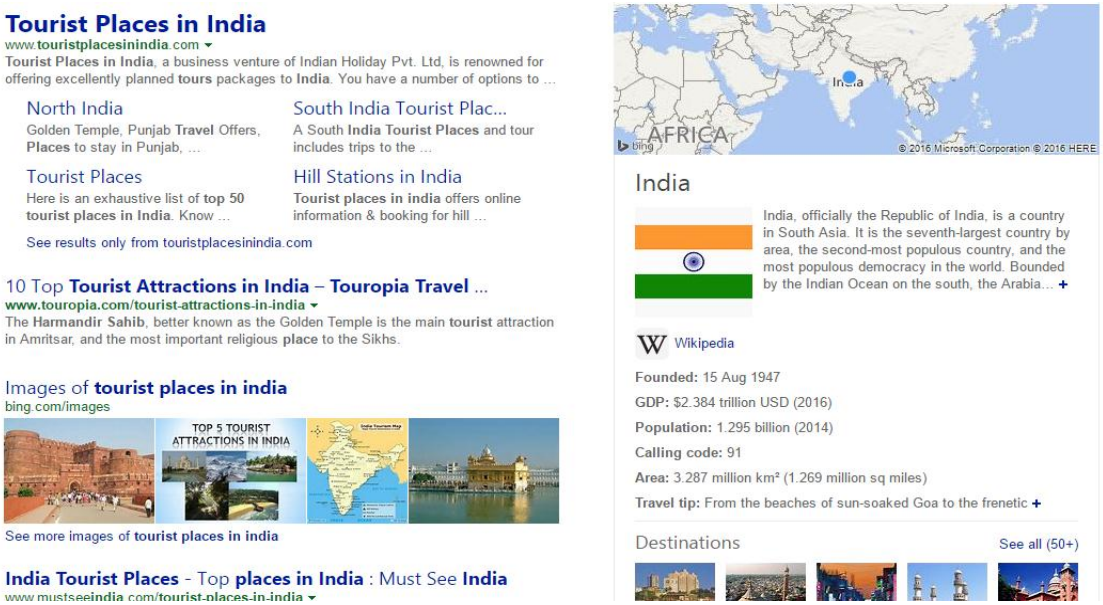
[2]Malabar Institute of Technology, Anjarakandy.
E-mail: neethu.t.regi@gmail.com

Abstract

The query facets are multiple groups of words or clauses that encapsulate the content covered by a query. When a query is given then from the search results the lists are extracted using the list extraction algorithm, all extracted lists are given weights, uniportant or noisy lists that occasionally occurs in a page are assigned by low weights. Similar lists are grouped into clusters using the weighted Quality Threshold Algorithm. Facets are then evaluated and ranked. High rank is given for lists extracted from similar context and lists having higher weight. We address the problem of finding query facets and the navigation to the desired web page. In order to increase the quality of the extracted list. We introduce a website wrapper called Anchor based data extraction system. The facets are generated based on the search interest of the user. The experiment is performed real time in Bing using the Bing search API. From the generated facets users can go to the desired high ranking pages by selecting the item in the facets. This can be used to generate facets in Bing.

INTRODUCTION

We address the problem of finding query facets and to link to the intended web page. A query facet is a group of items which outline and encapsulate one prime aspect of a query. Here a facet item is typically a word or a clause. A query may have many facets that encapsulate the information about the query from different aspects. We use the data mining methods for creating facets and to link to the intended web page.



Query facets provide fascinating and useful knowledge about a query and thus can be used to upgrade search experiences. Users can perceive some prime aspects of a query without browsing tens of pages. Query facets may provide direct information or instant answers that users are seeking. The facets are generated based on the users taste. The user can also go to the desired web page by selecting the item from the facets. Thus it saves the user's time wasted on searching for the data in tens or thousands of pages. We observe that prime slice of information about a query are usually presented in list styles and repeated many times among top retrieved documents. When a query is top k results are retrieved from the search engine. Then collect all documents to form a set of inputs. Query facets are mined by using four methods. List extraction, list weighting, list clustering, facet and item ranking. Compared to the previous work our work is using a website wrapper called Anchor based data extraction system for extracting high quality lists. These extracted lists are given weights by using the document matching weight and average invert document frequency. In the clustering the similar lists are then grouped using the weighted quality threshold algorithm. In the ranking step the lists are given higher ranks if it is extracted from unique context and having higher weight.

RELATED WORK

Luo et al. proposed Application of Internet Technology and Web Information extraction wrapper based on DOM for Agricultural Data Acquisition [3].It is the method of Web Information extraction wrapper based on DOM. Combining X-Path and pattern matching, it can deal with the two type of information at the same time under the guide of source and target knowledge library. Information extraction method is actually a text processing method. Rauch et al. proposed Knowminer Search - a Multi-Visualisation Collaborative Approach to Search Result

Analysis[4].Since the information provided on the internet is large It becomes difficult for the user to get apt information. Here faceted search interface provides the possibility to coherantly reduce the search result set. Friedrich et al. proposed Utilizing Query Facets for Search Result Navigation[5]. Facets provide a way to examine and go through the search result space. Features that rank facets based on their usefulnes to partition the search result documents. A very successful idea to generate facets for HTML documents is based on the extraction of lists from HTML pages. Simonini et al. proposed Big Data Exploration with Faceted Browsing[6]. Big data analysis now manage nearly every point of modern society. One of the most valuable means through which to make meaning of big data, and thus make it more helpful to most people, is data visualization. The faceted search allows the user to detail a query progressively, seeing the effect of each choice inside one facet on the available choices in other facets.

EXISTING SYSTEM

A query facet is a set of items which outline and condense one prime feature of a query. A facet item is typically a word or a clause. A query may have multiple facets that summarize the information about the query from different perspectives. [1] In existing system ie QDMiner, when a query q is given, we retrieve the top K results from a search engine and fetch all documents to form a set R . Then, query facets are mined by the following four steps:

1) List and Context Extraction: Lists and their context are extracted from each document in three formats:
2) Free text patterns –TEXTS and TEXTP:- extract items separated by , and extract items separated by:,—,-
3) HTML tag patterns - HTMLTAG :-SELECT For the SELECT tag, we simply extract all text from their child tags. UL/OL For these two tags, we also simply extract text within their child tags. TABLE We extract one list from each column or each row.
4) Repeat region patterns - REGION:-detect repeat regions in web pages based on vision-based DOM trees. Then extract all leaf HTML nodes.
5) List Weighting: All extracted lists are weighted, and thus some trivial lists that sometimes occurs in a page, can be given low weights.
6) List Clustering: homogeneous lists are grouped together to compose a cluster.
7) Facet and Item Ranking: Facets and their items are evaluated and ranked .

PROBLEM DEFINITION

In the existing system the list extraction method extracts low quality lists. If there is any change in the internal structure of the web page such as the page link or the addition of items to the website may produce low quality lists.

PROPOSED SYSTEM

In the proposed system website wrappers are introduced to extract high quality lists from trusted websites. A wrapper is a software that turns a Web source into a place that can be queried as a database. Website wrappers are used to extract structured information from a data repository, objects and relation. Here we used a website wrapper called Anchor Based Data Extraction System. A website wrapper can tolerate any change in the internal structure of the web page like the changes made by developers, page link. Website wrapper can be adapted to similar paper to gather and extract sources from web pages [2]. A website wrapper is created by identifying anchors on the typical web page. Then it is saved as a file for extracting data from similar pages on the website. An anchor is a textual element that marks the start or end of a data region or as a keyword in a data region that distinguishes it from the rest of the pages, such as the title, highlighted words, constants, keywords. For creating anchors the sample page is loaded into the embedded web browser. The elements of the page gets highlighted and the anchor is created based on the elements features. An anchor can be named based on the Id or class name. The region patterns are created based on one or more of the anchors. These act as the input to the data extraction algorithm. Where the lists are extracted and the data is stored in the XML format. After the list extraction the extracted lists are given weight, similar lists are clustered using the weighted quality threshold algorithm and then after clustering the facets are ranked, higher rank is given to the list extracted from unique context and list having higher weight. The facets are generated from the real time from Bing using the Bing search API. The facets are generated based on the user search interest. By keeping the search history of the user in a database on the server side. The user can also go to the desired website for the detailed information regarding the items in the generated facets.

CONCLUSION

This work presents a method of generating meaningful facets from the user query search results. The facets here are generated using four steps. List extraction, list weighting, list clustering and list ranking. Here the list extraction performed using a website wrapper called Anchor based data extraction system and the patterns are generated for these anchors. Then from these generated patterns lists are extracted using the list extraction algorithm. This method can extract high quality lists from the top k query search results. Hence these 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.

REFERENCES

[1] Zhicheng Dou, 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, 28(2):385–397, 2016.

[2] Ahmad Pouramini and Shahram Nasiri. Web data extraction using textual anchors. In 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), pages 1124–1129. IEEE, 2015.

[3] Liming Luo,Wen Lu, BingWei, Ye Qin, and YeQing Xiong. Application of internet technology and web information extraction wrapper based on dom for agricultural data acquisition. In Network and Information Systems for Computers (ICNISC), 2015 International Conference on, pages 327–331. IEEE, 2015.

[4] Manuela Rauch, Werner Klieber, Ralph Wozelka, Santokh Singh, and Vedran Sabol. Knowminer search-a multi-visualisation collaborative approach to search result analysis. In 2015 19th International Conference on Information Visualisation, pages 379–385. IEEE, 2015.

[5] Jan Friedrich, Christoph Lindemann, and Michael Petrifke. Utilizing query facets for search result navigation. In 2015 26th International Workshop on Database and Expert Systems Applications (DEXA), pages 271–275. IEEE, 2015.

[6] Giovanni Simonini and Song Zhu. Big data exploration with faceted browsing. In High Performance Computing & Simulation (HPCS), 2015 International Conference on, pages 541–544. IEEE, 2015.