COMPARISON: QT (QUALITY THRESHOLD) AND BATCH STS ALGORITHM FOR FACETS GENERATION

[1]Anju G R, [2] Karthik M
[1] Pg Student, [2]Asst.Professor
[1] Mohandas Collage Of engineering,
Mohandas Collage Of engineering, Trivandrum, India

Abstract— Searching is a charmed experience. There are many search engines tough it suffers many issues. The best search engine should be like, say for a nontechnical user to understand easily. One such method is the faceted search. Faceted search help searching by giving suggestions as facets. Facets are something that summaries an important aspect of a query. For example for the query computer the facets will be operating system, price, hard disk etc. So if a user focuses on hard disk then the user can go with the hard disk facet rather than searching for the computer keyword only. This is how facets help users. Facets help users by offering drill down option as a complement to the keyword input box. There are many methods to generate facets from the search results. In this paper compare two methods for facets generation from the search results. Mainly compare two clustering methods during the facets generation such as QT algorithm and Batch STS algorithm. Experimental results show that Batch STS is better than QT algorithm.

Index Terms—Facets, Search

I. INTRODUCTION

Searching is a very interesting and useful medium to gain knowledge. Search engines are used to gain the information. There are many search engines available today such as yahoo, Bing, Ask.com, Baidu, Wolframalpha etc. But still go through main problems say example information overload problem, giving unwanted information etc. So to avoid such problems faceted search is used. Searching is based on facets that are the faceted search. Facets are something which defines the important aspect of a query. For example query "Apple" the facets will be color, fruit, and laptop etc.

II. FACET GENERATION

The facets are generated using 4 methods list extraction from one of the famous search engines. From the search engine top 100 results are obtained. From the 100 results list are extracted from each document. The list is extracted using free text patterns, HTML tags and repeated region patterns. After getting the list the unwanted items are removed using weight listing, method. Then the list is clustered to get the facets. Here two clustering method are compared such as QT and Batch STS algorithm. The facets are ranked according to the frequency count then the facets are displayed to the user. Figure 1 shows the faceted search of an application. In the figure 1 can see the facets electronics, television also facets refined using price tags.

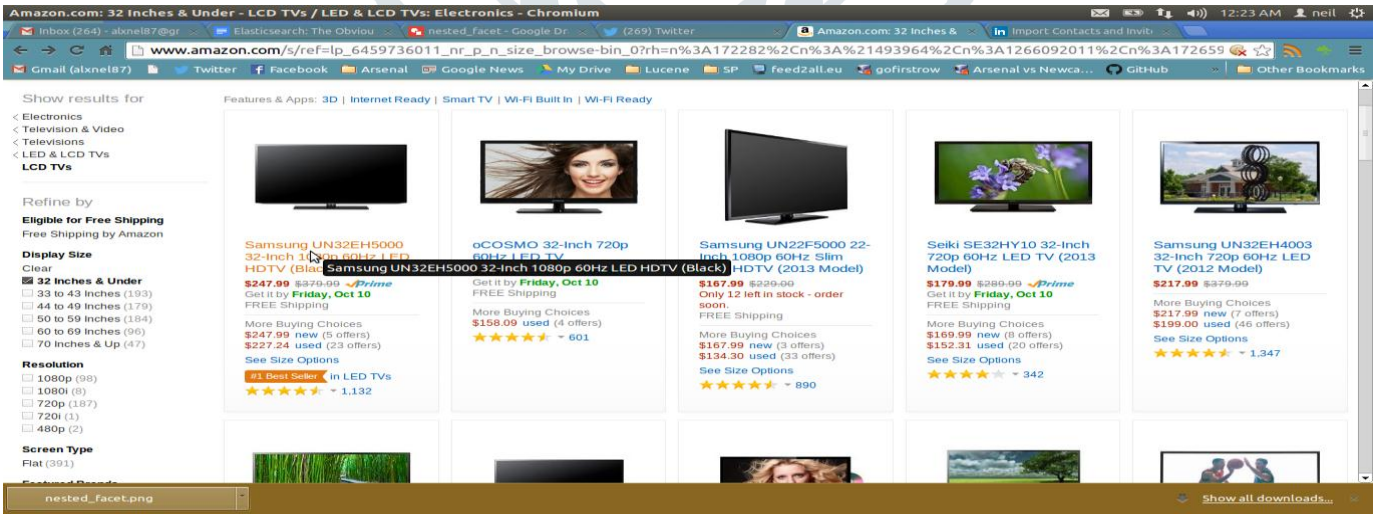

Figure 1: Example of faceted search application

III. ADVANTAGES OF USING FACETED SEARCH

- Ease of use
- Non-technical users can also understand the concepts of a particular query
- It gives a drill down options as a compliment to the keyword query
- Filtering of media types
- Fast indexing context
- Faster searching through structured filtering

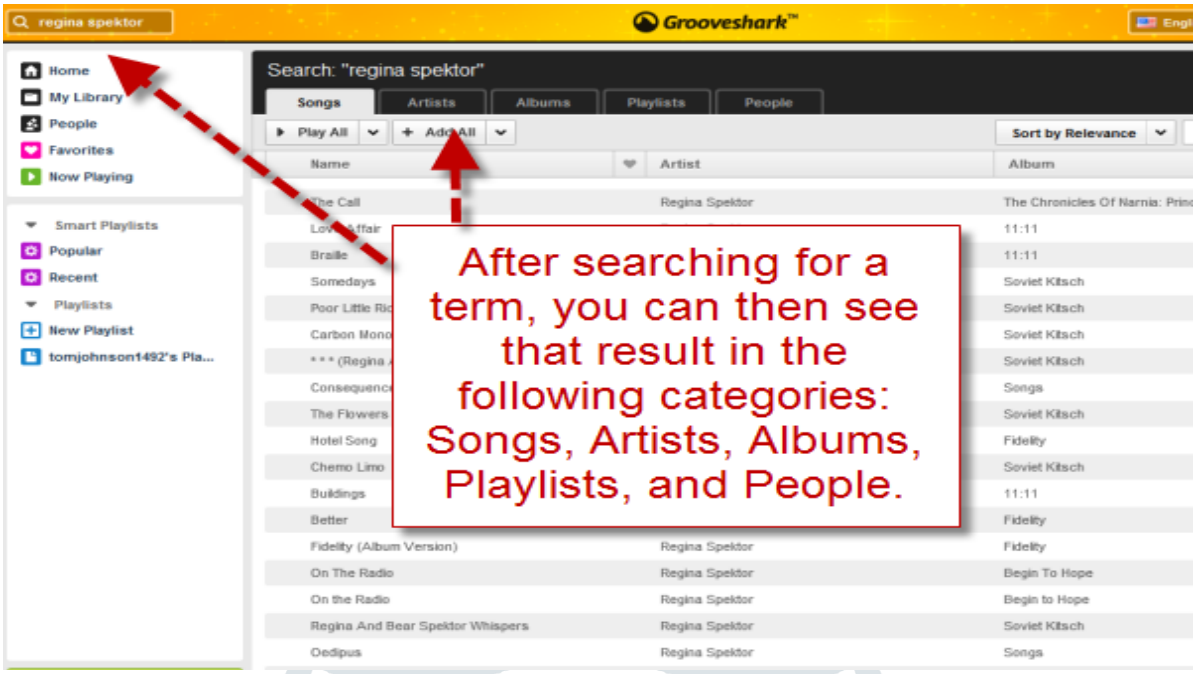

Figure 2: How faceted search is performed

The figure 2 shows how the faceted search is performed. When a user searches "regina spektor" can see some results such as songs, artists, albums etc. These are the facets, when the user select one of them the search results will be refined to those particular items only. These are the main benefit of using facets.

IV. QUALITY THRESHOLD ALGORITHM (QT)

Quality Threshold with Weighted data points is described as follows. [1]

- ✓ Choose a maximum diameter Diameter and minimum weight for the cluster
- ✓ Build a candidate cluster for the most important point by iteratively including the point that is closest to the group, until the diameter of the cluster surpasses the threshold Diamax. Here the most important point is the list which has the highest weight.
- ✓ Save the candidate cluster if the total weight of its is cluster is not smaller than user defined threshold
- ✓ Recurs with the reduced set of points.

In this paper, the weight of a cluster is computed based on the number of websites from which its lists are extracted.

V. BATCH STS ALGORITHM

The Batch STS algorithm is as follows: [2]

Algorithm BatchSTS

Input: S: the list set
θ r: the radius threshold
Output: top-k clusters which have top-k most lists

1. Initialize C = Ø;
2. For each element v i of S
3. If there exists any cluster C j where dis (vi, Cj) is not infinite
   4. Add v i into anyone of these clusters;
      5. Else
         6. Form a new cluster C new with the list $v\,i$;
         7. C = C ∪ C new;
8. For each non-single-point element C i of C
9. While the radius of C i is larger than or equal to θ r
10. For each list v j in C i
11. If dis $(v\,j, Ci) \geqq \theta$ r
    12. Exclude v j from C i;
    13. Check whether v j can be merged with other excluded lists;
    14. Output top-k clusters in C which has top-k most lists;
End

VI. COMPARISON

The analysis for clustering process is done using 2 methods.

- ✓ F measure
- ✓ Statistical approach

F Measure

F measure is used to calculate the accuracy of a clustering process. F measure is depending to precision and recall. The accuracy is high when F reaches 1.The F measure for QT algorithm is shown in table 1.

Table 1: F measure for BATCH STS

Query	Computer	Mobile	Games of thrones	Motorola	Season lost
True positive	6	2	7	5	10
False positive	2	1	3	5	4
precision	1	1	8	5	8
Recall	0.8571428 5	0.6666 666	0.46666 66	0.5	0.555 55
F	**0.8**	**0.6666 666**	**0.56**	**0.55555**	**0.625**

The batch STS algorithm the F measure for keyword computer is 0.8, for mobile 0.666 and so on as in the above table 1.

Table 2: F Measure for QT

Query	Computer	Mobile	Games of thrones	Motorola	Season lost
True positive	6	2	7	5	10
False positive	3	4	5	4	6
precision	1	1	8	5	8
Recall	0.85714 28	0.6666 666	0.4666 666	0.5	0.5555 55
F	0.75	0.4444	0.5185	0.52631	0.5882

The QT algorithm the F measure for keyword computer is 0.75, for mobile 0.444and etc as in the above table 2. From this it's clear that **Batch STS algorithm is better than QT**

Statistical Approach

Statistics about the Query Facets Generated with Search Results. To better understand the quality of the generated facets, show some statistics about the generated query facets on both algorithms. From the table it's clear that facets per query for QT is 6.8 (ratio) and for Batch STS it is 7.0. Like that list per facets for QT is 2.764 and for Batch STS it is 3.3 from this also its understandable that Batch STS is far better than QT.

Table 3: statistical approach

	QT	BatchSTS
Queries	5	5
Results per query	100.0	100.0
Lists per document	10.987	10.987
Items per list	7.9633802816901	7.9633802816901
Facets per query	6.8	7.0
Lists per facets	2.764	3.3

V. CONCLUSION

Faceted search help users by giving drill down option as complement to keyword input box. Faceted search is implemented in many applications such as "eBay, Amazon" etc. However when it comes to the general web is difficult to obtain facets due to the heterogeneous nature of web. From the general web the facets are mined using many steps. In this paper compare two algorithms for the clustering process such as QT and Batch STS. From this can see that BATCH STS is better than QT for clustering the list to get facets.

REFERENCES

[1] Automatically Mining Facets for Queries from Their Search Results" Zhicheng Dou , Member, IEEE, Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song. IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 2, February 2016

[2] "IncreSTS: Towards Real-Time Incremental Short Text Summarization on Comment Streams from Social Network Services". Cheng-Ying Liu, Chi-Yao Tseng, Ming-Syan Chen, Fellow, IEEE DOI 10.1109/TKDE.2015.2405553, IEEE Transactions on Knowledge and Data Engineering

[3] C. D. Manning, P. Raghavan, and H. Schtze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008