International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-6 Issue-5, June 2017

Dynamically Building Facets from Their Search Results

Anju G. R, Karthik M.

Abstract: ***People are very passionate in searching new things and gaining new knowledge. They usually prefer search engines to get the results. Search engines become an important way to get the information. But many search engines fail to give some request to the users since there are same words which have different meaning such as apple, say it's a fruit, mobile, laptop. So if there is ranking based on these, the searching will be a pleasing experience's. There are some methods for these such as searching based on facets. There are some exiting methods to gain facets from the search results and display the facets such that the user can select corresponding facets. Then the search results will be refined to those particular facets only. In this paper mainly focus on those facets that mean after the facets generation, the facets will be checked before displaying to the user. There are some facets such as "women watch, women's watch " , "Season one, season 1" these two have same meaning so before displaying the facets these similarities should be checked and only one facets should be displayed. Part of speech is also checked. Experimental results shows that checking these type similarities improve the facets thus it can improve the searching experiences in many ways.***

Index Terms: Faceted search, Facets, Intent

I. INTRODUCTION

Information gathering is a very useful in every one's life. People are eagerly waiting to get the information's. There are many search engines available today; most search engines have abundant data's.

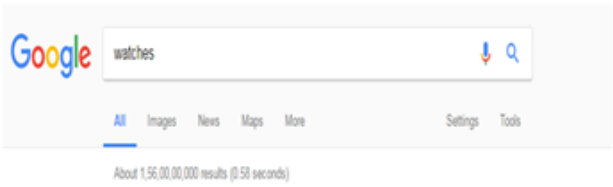


Figure 1: Google Search Engine

For example in the figure 1 can see that when a user type "watch" as a keyword there are up to 1,56,00,00,00 documents or results available in 0.56 seconds. As a user it is very difficult to process all these documents and get the appropriate results. Faceted search become an easiness solution for these. Faceted search is a prior to clarify the search results using facets. Facets are something that digests a meaningful aspect of a query. For example for the query laptop the facets will be configuration, price etc.

Manuscript published on 30 June 2017.
***** Correspondence Author (s)
Anju G. R, Department of Computer Science, Mohandas College of Engineering and Technology, Thiruvananthapuram (Kerala)-695544, India.
Karthik M, Department of Computer Science, Mohandas College of Engineering and Technology, Thiruvananthapuram (Kerala)-695544, India.

Figure 2: Facets for the Query Watch

Figure 2 shows the facets for the query watch. When a user types a query "watch" in the search engine say Google, the facets will be gender, brand, supporting feature. For the brand facets the facet terms will be Cartier, omega etc. Facet terms will be a word or a phrase. When clicking or selecting a particular facet term the corresponding result only is visible. Faceted search is in a job with many applications such as: eBay, Amazon, twitter" etc. In this application the facets are generated to the integral corpus manually or semi automatically. But this method is difficult in genaral web due to the diverse nature of the web. Diverse means due to the vast and dissimilar content in the web it is difficult to create facets from results.

To create a query depended facet generation is a headache problem. There are many methods to generate facets manually or semi automatically such as product application sites but facet generation from the search results has only some existing system. That means there are only few works based on facets generation from the search results.

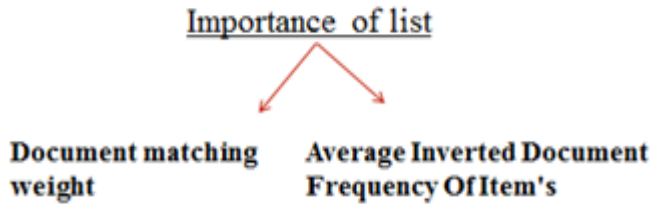
II. EXISTING SYSTEM

In the existing system for mining query facets there are mainly two stages. [1] In this approach it naturally cumulates recurrent items in the top search results. That is first of all from a search engine extract top results. From the top results facets are generated. Figure 3 shows the flow chart for the existing system. First a user inputs a query i.e. keyword. For the particular keyword from the search engine documents are retrieved. The top results are fetching from a search engine .In the existing system it uses 100 results from the search engine. So the search results contain all the fetched documents. From the fetched documents list are extracted. The extracted list may contain nonessential items. So that should be avoided to get the good results. The list contains all the extracted list .List contains unwanted items that should be removed. For that list Refining is used. Refining consist of two sections.

- ✓ List weighting
- ✓ List clustering

In the list weighting the corresponding list is weighted according to the frequency count. It is based on two factors [1].

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)



By combining the above two factors, will get the weight for the list. [1] Next section is the list clustering section. In this section list of similar items will be grouped to form a facet. Clustering is done using two methods in the existing systems

- ✓ Quality Threshold algorithm [1]
- ✓ Batch STS algorithm [2]

Both these methods are used for the clustering process. From these Batch STS performed better than QT [3]. After the refining section facets will be generated. These facets are ranked according to the frequency count and return back to the user.

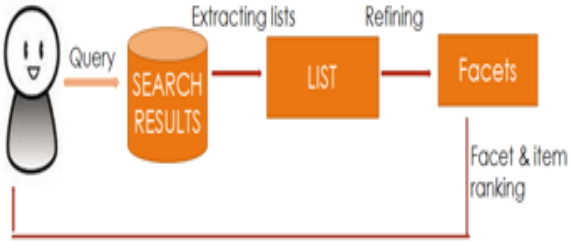


Figure 3: Flow of work

Figure 4: Existing system

Figure 4 shows different steps in the existing system. For the list extraction it uses three methods

- ✓ FREE TEXT PATTERN
- ✓ HTML TAG PATTERN
- ✓ REPEATED REGION PATTERN

The free text means the paragraph's lies in the websites. So first of all need to identify the p tag in the web page html design. Extract the text within the p tag. Split the paragraph into sentences using full stop. Apply the lexical pattern to each sentence to obtain the corresponding list. Then in the extracted list refining is performed to get the facets. The facets and items are ranked according to the frequency count. Finally it is returned back to the user.

III. PROPOSED SYSTEM

Figure 5: Proposed System

The proposed system is shown in figure 5. There are mainly four modifications are there. First modification is in the document fetching section. Here use search results from Google and yahoo To extact the list. By combaing the search results the facet qualtiy can be increased. Second modification in the list extraction area such that strong patterns will be held in another list. Third modification is done in the facet and item ranking. In the existing system after getting the facet the facets and items are ranked according to the frequncy count and it is returned back to the user. In the proposed system facet simmilarites are also checked such as "season1" and "seasonone" have the same meaning. Such simmilarities are also checked before displaying to the user. Part of speech are also checked.

A. List Extraction

Figure 6: List Extraction

In the list extraction main focus is on the HTML TAG. Figure 6 shows the new approach in list extraction. In the existing system list is extracted from select, ul, ol tags. The select, ul, ol tags will be in one list. But as in the figure 7 inside the "ul "or "ol" tags strong contents means bold contents will be there. These strong contents will be placed in another list such that it will affect the quality of facets. Because assuming that the important things of a query are represented in bold or strong patterns.

Figure 7: Strong Tags

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In the figure 7 memory, mass storage etc. are represented in ol tags. In the ordered list some are represented in strong patterns. These strong patterns are put in one list rather than only ul and ol list.

B. Facet And Item Ranking

This section is the most important. Before displaying to the user the part of speech is removed and some resemblance are also checked such as "women's watch" and "women watch". This type closeness is also detected. Figure 8 shows the part of speech. These parts of speech are removed to get the good facets. When removing these parts of speech can improve the quality of facets since these are not representing any important things of a query.

Figure 8: Part of Speech

IV. EXPERIMENATL RESULTS

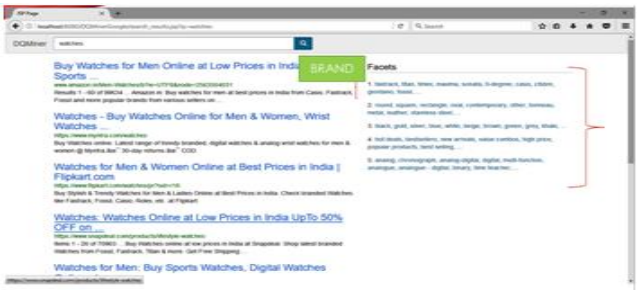


Figure 9: Searching

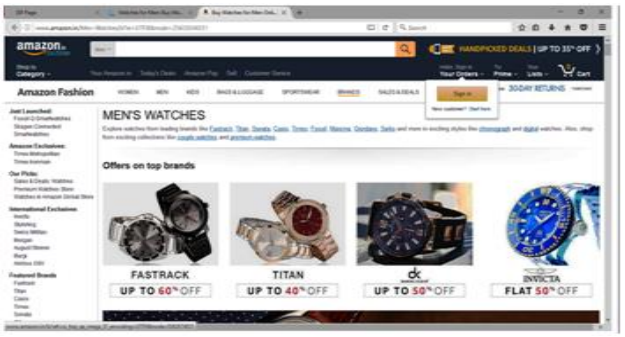


Figure 10: Displayed URL

The figure 9 and 10 shows the experimental results. Figure 8 shows when a user inputs a query for example "watches" first of all fetch documents from Google and yahoo results. From the results facets are generated. The corresponding facets are displayed. When type watches keyword the first facet is the brand as shown in the figure 9. For the brand facet the facet items will be fast track, titan etc.

In the figure 10 when a user selects one of the facet the corresponding top results will be opened as a new tab. The new tab contains the sites which have highest rank.

V. EVALUTIONMETHODLOGY

Ranking effectiveness of facets [1]: Aim to rank good facets before bad facets when multiple facets are found. As have multi-level ratings, adopt the nDCG measure (Normalized Discounted Cumulative Gain), which is widely used in information retrieval, to evaluate the ranking of query facets.

fp-n DCG [1] - purity aware nDCG based on the first appearance of each class. Different from the original nDCG, further consider the purity of each facet ci by multiplying [1] DGi bythe percentage of correctly assigned items. [1] [7] further use the evaluation metrics PRF and wPRF proposed by Kong and Allan [4]. PRF is a harmonic mean of precision of facet terms, recall of facet terms and facet clustering F1. wPRF further takes into account the ratings associated with query facets, and it uses weighted facet term precision, recall, and clustering F1. Neither PRF nor w PRF accounts for facet ranking effectiveness Discounted cumulative gain [5] (DCG) is a measure of ranking quality. In information retrieval, it is often used to measure effectiveness of web search engine algorithms or related applications. Using a graded relevance scale of documents in a search-engine result set, DCG measures the usefulness, or gain, of a document based on its position in the result list. The gain is accumulated from the top of the result list to the bottom, with the gain of each result discounted at lower ranks.

[5] Two assumptions are made in using DCG and its related measures.

- ✓ Highly relevant documents are more useful when appearing earlier in a search engine result list (have higher ranks)
- ✓ Highly relevant documents are more useful than marginally relevant documents, which are in turn more useful than non-relevant documents.

DCG originates from an earlier, more primitive, measure called Cumulative Gain.

List extraction: For the list extraction use the purity measure.

VI. PERFORMANCE ANALYSIS

A. List Extraction

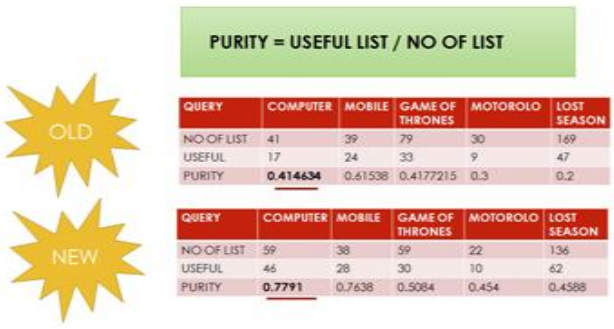


Figure 11: Purity Measure

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

For the list extraction performance analysis is calculated using purity measure. Purity measure is used to calculate the accuracy of a method. In the figure 11, purity for new and old method is calculated. Checked for 5 queries, such as computer, mobile, games of thrones etc. In 5 queries the new method is better than the old method. For the computer keyword the purity when using old method is 0.414634 and for the same keyword computer when using the new method the purity is 0.7791. This shows the improvement when using the new method.

$$Purity = \frac{no\ of\ corrected\ items}{total\ no\ of\ list} \quad (1)$$

B. Facet Ranking

Ranking effectiveness of facets is calculated using two methods. n-DCG and fpn-DCG.

OLD

QUERY	COMPUTER	MOBILE	GAMES OF THRONES	MOTOROLA	LOST SEASON
DCG	9.49	12.29	13.74	6.88	18.22
IDCG	12.91	12.29	16.65	7.95	24.11
nDCG	0.73508907	1	0.82522	0.8654	0.7557

NEW

QUERY	COMPUTER	MOBILE	GAMES OF THRONES	MOTOROLA	LOST SESON
DCG	11.36	12.29	13.86	6.88	18.47
IDCG	12.91	12.29	16.65	7.95	24.11
nDCG	0.8799380	1	0.83243	0.8640	0.7660

Figure 12: Normalized Discounted Cumulative Gain

OLD

QUERY	COMPUTER	MOBILE	GAMES OF THRONES	MOTOROLA	LOST SEASON
DCG	9.12	12.01	13.47	6.43	18.03
IDCG	12.91	12.29	16.65	7.95	24.11
nDCG	0.7064291	0.9772	0.8090	0.80880	0.7478

NEW

QUERY	COMPUTER	MOBILE	GAMES OF THRONES	MOTOROLA	LOST SESON
DCG	11.01	12.01	13.59	6.43	18.2
IDCG	12.91	12.29	16.65	7.95	24.11
Fpn-DCG	0.852873	0.97721	0.8162	0.80880	0.7548

Figure 13: fpn-DCG

In the figure 12 normalized discounted cumulative gain for both new and old approach is tested. In the old method for the computer keyword n-DCG value is 0.73508907 and for the new method n-DCG value is 0.8799380. From this it is clear that new method is far better than new one.

In the figure 13 shows the purity aware first appearance n-DCG. Here also can see the improvement in the new method.

C. Search Results

NO OF RESULTS	GOOGLE	YAHOO	GOOGLE & YAHOO
10	0.8	0.75	0.7777
20	0.916	0.648	0.85

Figure 14: Search Results Quantity

In this further analyze with number of experimental results. For this use the purity measure. Experimental results are shown in Fig. 14. This figure shows that the number of results does affect the quality of facets. Query facets become better if more search results are used. This is because more results contain more lists and can generate more facets.

More results also provide more evidence for voting the importance of lists, hence can improve the quality of facets. But when using the yahoo results only the facets getting are not good as compared to the Google and combined Google and yahoo search. In the figure 15 it's clear that when using yahoo results only as the no of search results increased the quality of facets are decreased.

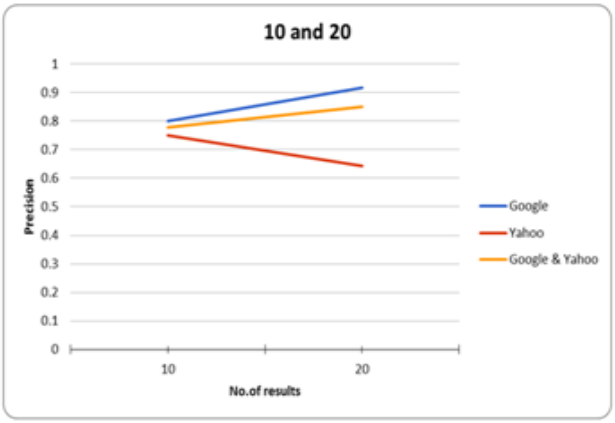


Figure 15: Graphical Repression of Search Result Quantity

VII. CONCLUSION AND FUTURE WORK

People are very eager to gain new things. In this paper focusing about facted search. The main advantge of facted search is that even for the non –technical user can also use the search engine. This facted search make searching task simpler and very easy. In this paper manily focusing on facet generation. Propose a new method for extacting list from the search resuls and find a new method in the facet and item ranking.

In the list extraction a separate list is considered for the strong format in the HTML web page design rather than only list for ul and ol tags as in the existing system. Experimental results shows that by considering the strong patterns as another list can improve the quality of facets.

In the existing systen aftr getting the facets ,the facet and items are ranked according to the frequency count and it is displayed to the user. But in this approach before giving facets to the user the facets adjacency are checked. Such as "womens watch and women watch have nearness meaning. So such nearness are also checked. Expermiantal results shows that when ading these can improve the quality of factes. Part of speech are also checked. The most common parts of speech are nouns, pronouns, verbs, adjectives, adverbs, conjunctions, and prepositions. These part of speech are also removed . Thus can improve the facet quality.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

This can be further improved using many aspects such as some semi supervised bootstrapping list extraction algorithms can be used to iteratively extract more lists from the top results . Machine learning can be implemented to give automatic description to the facets such as barnd, colour etc.

REFERENCES

1. "Automatically Mining Facets for Queries from Their Search Results" IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 2, February 2016 Zhicheng Dou, Member, IEEE, Zhengbao Jiang, Sha Hu, Ji-Rong Wen, and Ruihua Song
2. "Facets Mining From Search Results Using BatchSTS Algorithms" in IJARTET, Volume 4, Special Issue 6, April 2017, Anju G R, Karthik M
3. "Comparison: QT (Quality Threshold) And Batch STS Algorithm For Facets Generation" JETIR (ISSN-2349-5162) April 2017, Volume 4, Issue 04, Anju G R, Karthik M
4. 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.
5. https://en.wikipedia.org/wiki/Discounted_cumulative_gain Google Wikipedia
6. 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.
7. W. Kong and J. Allan, "Extending faceted search to the general web," in Proc.ACMInt. Conf. Inf. Knowl. Manage., 2014, pp. 839–848.
8. "Mining Queries From Search Results : A Survey" Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-12, 2016, Anju G R, Karthik M

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)