Facet Mining from Annotated Web Pages

Sijin P[1], Champa H N[2]
[1,2]*Department of Computer Science and Engineering, University Visvesvaraya College of Engineering*

Abstract **- The query dependent multidimensional views of a search query which describe and summarize some important aspects of search are called facets. The Faceted search allows the users to search on a facet list of a query to pick out the desired one without browsing for a long time. The proposed facet mining framework called Facet Mining from Annotated Documents (FMAD) integrates keyword search by category browsing and produces an interface which has several conceptual dimensions. The FMAD is designed to achieve an interactive data summarization process by liberalizing topic identification and interpretation on user side called faceted search. The proposed ontology based query dependent facet framework follows a series of phases from initial facet weighting to facet item ranking in order to produce high quality clusters. The proposed Quality Threshold with Weighted Data Point and Web Page Count (WQTWC) algorithm is used to produce group of items which are similar in same group clusters and differ with other group clusters. In FMAD facets are automatically extracted from document description, annotated documents, and metadata records.**

Index Terms **- Item ranking, Facet, Data Point, Facet quality.**

I.INTRODUCTION

A query may have different perspectives while searching over an open domain data repository. Each of these perspectives can be represented with different facets. Each facet is a set of items which describe and summarize one important aspect of a query. User can display the query facets along with top search results, since facets reveals the different aspects of a query, user will get an additional set of results which may related to the given query and could be used for approximate search process [1], [2]. In facet-oriented search user intention is diverted to the required facet just a few keys away without much overhead on query evaluation processes. A keyword search is recommended for a faceted search if it has the following characteristics [3], answers are unable to fetch on initial iterations, the user is able to interact with the results obtained and could reformulate the query, and in multiple item searches.

The quality of a facet can be measured in terms of specificity and dispersion. Specificity is the quality of belonging or relating uniquely to a particular subject. Dispersion can be defined as the action of distributing something or appearance of a facet on multiple lists. The major advantages of faceted search are nothing but without browsing for a long-time user could view the facet lists of a query and could pick out the desired one. Direct and instant answer groups are available with faceted search.

e.g Cell phone groups with display size, color, OS used

Table 1. Query facets for query: "Watches".

Facet Name	Group Items
Watch category	Men, Women, Kids etc.
Watch brand	Alpina, Citizen, Favre Leuba, Movado, Victrinox etc.
Watch Color	Black, Brown, Pink, Red etc.
Watch Type	Analog, Digital, quartz etc.

Similar to normal and approximate query results facets are also abundant in nature. In order to avoid the facet boom it is possible to rank the facets and could display the top-k diversified link [4], [5], [6], [7]. Table 1. shows the facet lists for item "watch". Facet lists for watch is obtained on four categories namely Watch category, watch brand, Watch Color and Watch Type.

The proposed facet mining for Table 1. Data generates faceted items with high ranks by a systematic ranking process based on metadata and is given in Figure 1. Transitional query suggestion approaches such as annotations, query logs, online summarization methods are used to design database schemas, and to generate metadata attributes [8], [2]. In proposed system facets are designed to provide a list of choices to user, query suggestion and annotation is used to represent multi-view navigation over the links.

The facet mining process over web pages is comparatively complex because of the abundance data to process, cross references, and data duplication in web pages. If the system is going behind all the

conceptual dimensions for a facet it should be a cumbersome task. The FMAD performs a weighing process followed by ranking facets approach to sort out top-k lists which are specific in list wise and differ in cluster groups of facets.

II. LITERATURE REVIEW

Grouping of facet lists with same concepts can be used to relate them. These grouping include concepts and subjects and this in turn contain a hodge-podge of subject category [9], [10], [11]. Common aspects of exploratory tasks are uncertainty, ambiguity and discovery [3]. Faceted search integrates free text search with structured querying and facet count serves as a context for navigation in a discovery driven analysis of data [12].

The facet optimization process is used to list out facets that could effectively partition the product search space so that, the user can easily drill down and find out the desired product [13]. The correlated facets problem associated with basic facet paradigm can be reduced by providing a tree indexing scenario. In faceted search over the web with heterogeneous nature, a query dependent automatic facet is good because it generates facets for a query instead of the entire corpus [14], [15].

The facet dimensions are formed by attribute of target articles. By analyzing the induced sub graph of a Relevant Category Hierarchy (RCH) it is possible to identify individual facets which are occupied in various categories, similar to concept cluster vectors for query terms in Concept network. A faceted interface can be defined as an interface consists of k facets [16], [17].

Faceted metadata is also known as content-oriented category metadata because it contains an orthogonal set of categories. The metadata may be flat e.g Alpina or hierarchical Watch brand- Alpina or it may be single or multi-valued in nature [18], [19]. In a facet-oriented search user search is diverted to the selected facets and the user intention is preserved [5], [8].

Facet generation process uses both abstract and extract summaries especially in the case of metadata oriented faceted search. Predefined fixed categories of search results and tedious search reviews can be used for facet generation, but they produced insensible results because of the inability to capture differences in search results [7], [20].

III. METHODOLOGY

The proposed facet mining process has various phases such as facet extraction, facet weighing, facet list clustering, facet ranking, item ranking. Facets are used to prepare multiple groups of semantically related queries. These are more informative than traditional query suggestion methods and help users to fetch out better query more easily.

3.1. Facet extraction

In a web document facet for a query is available in different forms such as list, paragraphs, html tags, XML tags etc. which could be extracted by using query processing languages such as XQuery or other wild card-based programming languages with regular expression support. The proposed FMAD performs facet extraction from metadata of documents. It can also able to extract required information from web forms by applying tag level extraction. If enough information is not available from metadata forms then it used Fuzzy Conceptualization Model (FCM) to obtain relevant results.

The metadata of document contains lot of information about the contents and properties related to that document. Meta name keyword says about the nature of documents to which it is related. The metadata keywords like "wristwatch for men", "wrist watch online", "digital wrist watches" are pointed out that the document says about the wrist watch category and sales. The meta properties like title, type, url, image, description, site name, sign-in client-id, sign-in cookie policy, sign-in scope etc. are useful for achieving document categorization and data security. The following command is used to extract keyword "watches" from metadata list of documents in order to know about relevancy measure of documents on the context of term "watches". Usually web documents are annotated with relevant keywords and which are added to metaname tag of metadata representation pages. The metadata information extraction is criticized for its deficiency of entire document coverage with fuzzy set of query base terms. The FMAD is designed to use annotated documents in which document is annotated based on its structure and contents trained over a workload.

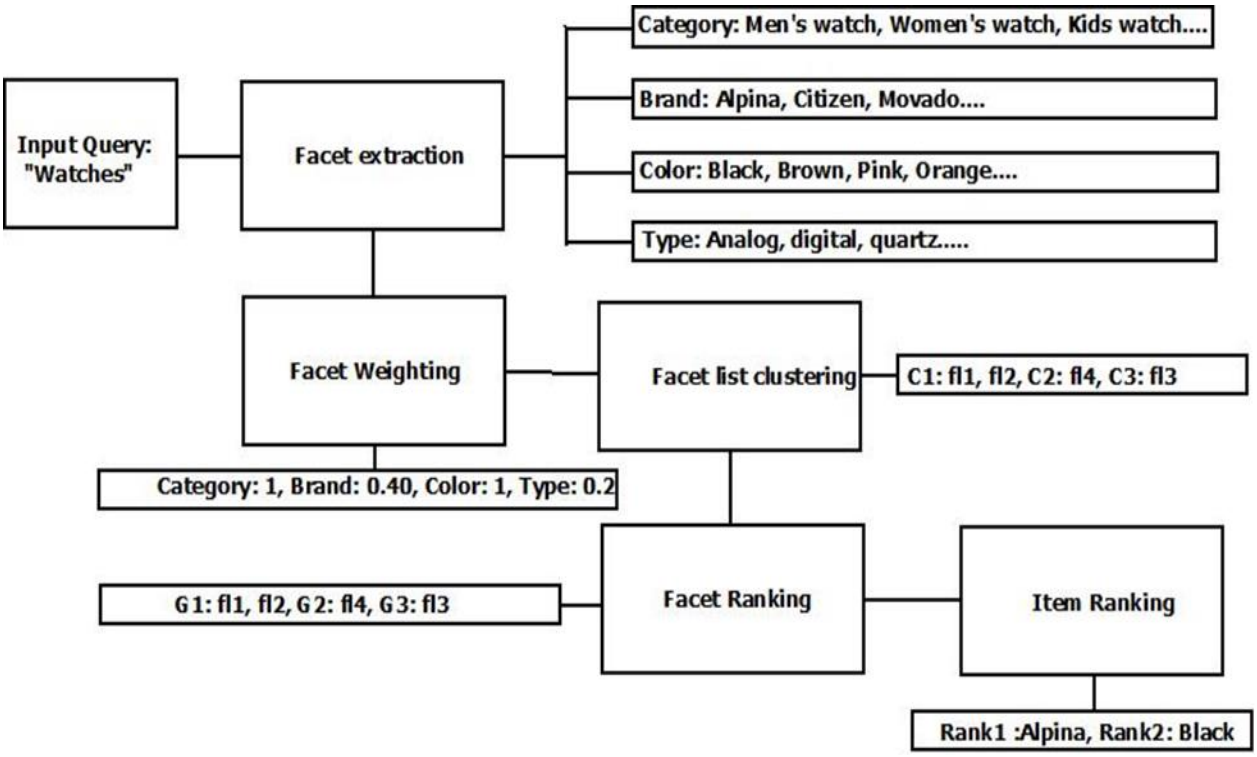

Fig. 1. Open domain facet mining process

In such an approach each document is categorized and localized in repositories and its ontology is well written in metadata more than mere document representation. A separate data adaptation insertion form is used for this purpose and used to upload from work sites. The form was filled by domain experts initially and automated adaptation is practicing now with trained workload and query logs [2], [21], [22]. The following command is used to obtain concept list (facet lists) for keyword "watches" from document metadata.

let *feeddoc* := *doc*("*ClueW eb:xml*")

for keyword at *i in feeddoc=ClueW eb* return *keyword=metaname*[*contains*(*., 'watch'*)]/*text*()

There are lots of web scraping tools such as Import.io, Octoparse, Uipath, Kimono etc. can be used to extract desired data from web pages. When using Java, c#, Python, PHP etc. user can program on to obtain desired results. The following command is used to obtain items from a list using a tag level extraction on html tag "li".

var x = document.getElementsByTagName("li"); document.getElementById("demo").innerHTML = x[0].innerHTML;document.getElementById("demo") .innerHTML = x[1].innerHTML;

Table 2. lists out the results for metadata extraction and tag based extraction process for < *li* > with java. After list extraction the obtained results listed as facets are Category, Brand, Color and Type.

Table 2. Examples for data extraction from web page.

Extraction Type	Facet group	Faceted data
Metadata	Watch brand	Shop watches from 25 + brands like
Metadata	Shopping	helios watch store
Metadata	Watch price	Shop by price-range
Metadata	Watch category	Men and women watches
Metadata	E-shopping	https:myntra.com
Metadata	Watch type	gold, titanium, ceramic
Tag based	Price range	price=1000 - 10000,

3.2. Facet Weighting

Facet weight in a document is calculated based on facet count in document and the number of shared keywords in facet keyword lists and document. Let Mk be the metadata keyword lists for a facet. If a document contains all these keywords or shows an FCM relation to the given keyword list then this document is related to the facet query. The strength of this relation is depended on the cardinality of shared keyword set or the value of Effective Assessment Coefficient (EAC) in case of FCM. If a document accounts for more number of a particular facet then that document is more belongs to that

particular group of facets so facet count is valuable for Facet Weight(Facetw) calculation. Inorder to balance the effect of common items in a document Inverted Document Frequency (idfe)of that
document can be considered during weight calculation and the entire equation is given in (1). ClueWeb dataset, a web scraped data collection is used for facet weighing demonstration.

$$Facet_{w} = Doc_{w} * Facet_{count} * \frac{idf_{e}}{|l|} \quad (1)$$

Where

$$Doc_{w} = \frac{N_{ld}}{|l|} \quad (2)$$

where Nld ≤ Mk is the number of shared keywords in facet keyword list and the document used and l is the total number of lists identified. idfe is calculated as

$$idf_{e} = log\frac{N - N_{e} + 0.5}{N_{e} + 0.5} \quad (3)$$

where N is the total number of documents in the given corpus and Ne is the number of documents which contain the given item e. After facet weighing normalized value list obtained is given as (Category: 1, Brand: 0.40, Color: 1.0, Type: 0.2).

3.3. Facet list clustering

Facet weight provides valued list for a search but facets are groups which drilled to a target item and differed each other among groups. A facet with multidimensions says about the various directions of search. During facet clustering list with similar intentions are grouped together. If two facet lists are related proportionally distance between them is calculated as in (2). Quality Threshold with Weighted data point with web page count (WQTWC) is used to solve this problem.

The pairwise distance obtained between the facet lists are d(fl1,fl2)=0.2, d(fl1,fl3) = 1, d(fl1,fl4)=0.94, d(fl2,fl3)= 1, d(fl2, fl4)= 0.91, d(fl3, fl4)= 1. WQTWC algorithm set a cluster diameter of user choice and here it is 0.96. All the points with threshold value up to 0.96 is locked in initial cluster as given in Algorithm 1. The Algorithm 1 chose minimum weighted list Wmin from faceted list. Then it chose a threshold diameter Diametermax and randomly selected a target facet list with weight Target weight. The pairwise distances among lists are calculated and made a core cluster for initiation of clustering by set Targetweight as locus of the core cluster. Then added all faceted list with diameter less than Diametermax to core cluster. The remaining faceted lists are also clustered similarly. Checked the lookup table of faceted list for getting website count and performed a reclustering operation to produce final clusters. The algorithm produced three clusters namely C1, C2, C3 and shown in Figure 2. The faceted lists d(fl1, fl2, fl4) are grouped in cluster C1. Faceted lists d(fl3) is grouped in cluster C2 and faceted lists d(fl4) is grouped in cluster C3. The typical WQT algoritm works on weighted list and grouped fl4 in cluster C1. The normalized website count for the given faceted lists is calculated as (Category: 1, Brand: 1, Color: 0.6, Type: 0.01). The website count for fl4 is comparatively less and it increased its threshold distance above 0.96 and it is grouped in C3. In this way

$$d(f_{11}, f_{12}) = 1 - \frac{|f_{11} \cap f_{12}|}{min(|f_{11}|, |f_{12}|)} \quad (4)$$

WQTWC produced high quality clusters. Each cluster produced relevant results and at the same time they are differed from other cluster groups.

3.4. Facet ranking

Facet ranking is the process of grouping lists based on their hamming distance in each group. Hamming distance between fingerprint of lists context is used to measure the similarity between two lists. Weight of a group is the number of lists in that group. A list

Algorithm 1: Quality threshold with weighted data point with web page count

Data: The facet keyword lists of various items.
Result: The Pairwise distances among faceted list and facet clusters for the query.
initialization;
1. Choose a minimum weight W_{min} for facets.
2. Choose a threshold diameter $Diameter_{max}$ for facets.
3. Choose a target facet $Target_{weight}$ with weight ≥ W_{min}
4. Calculate the pair wise distance among all pairs of facet lists according to (2).
5. Make a core cluster with $Target_{weight}$ as locus.
6. Add all faceted list with weight ≤ $Diameter_{max}$ to the core cluster.
7. Go to step 2 and Cluster the remaining faceted list according to their weight.
8. Calculate website count Web_{count} for each facet list.
9. Calculate the new weight by multiply(2) with Web_{count}.
10. Go to step 2 and restructure the clusters for new facets by feed backing current cluster details.

Algorithm 2: Facet Ranking Algorithm

Data: The facet clusters obtained from Algorithm 1.
Result: Group of similar lists inside clusters.
initialization;
1. Set the minimum weight W_{min}=1.
2. A duplicate threshold is set as FL_{dup}.
3. Choose the cluster with list weight is high and initiate a sub group.
4. Add further list to the sub group untill diameter of the cluster surpasses the threshold FL_{dup}.
5. Save the sub group and remove other lists. Make a core cluster with $Target_{weight}$ as locus.
6. Go to step 2 and do it for all other clusters.

grouping algorithm is used for this purpose. The Algorithm 2 sets minimum weight Wmin=1 and a threshold value FLdup is identified for the similarity hamming distance to form a subgroup in selected clusters. The cluster with highest weighted lists is identified and highest weighted list is made as the locus of subgroup within the group. The other lists with weight less than FLdup added to the new subgroup. The same process has done to other n topmost clusters. The hamming distance for Watch category (d(fl1) and Watch brand (d(fl2) is calculated as in (5) and the subgroup formed in cluster C1 is shown in Figure 3.

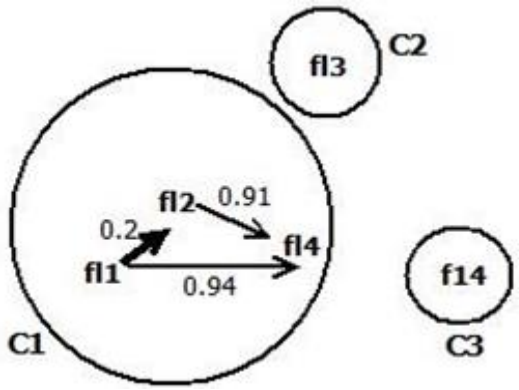

Fig. 2. Facet Clustering with WQTWC Algorithm

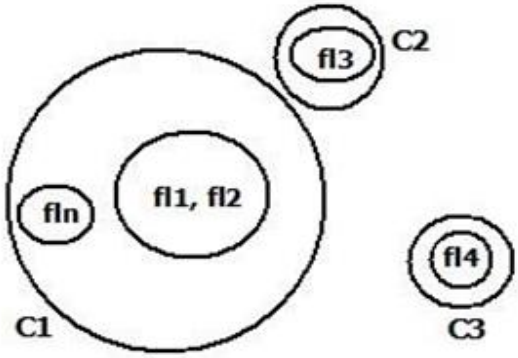

Fig. 3. Facet ranking in clusters

where LFP is the length of fingerprint used by default it is 64 and HDfl1, fl2 is the hamming distance between two faceted list (fl1, fl2).

3.5. Item ranking

The importance of an item e in a list depends on its rank in the list and the number of lists accessing it. The rank of facet in a group depends on facet item count. The weight of an item in a faceted list is calculated as in (6).

$$W_{\text{item}} = \sum_{\text{fl}_i \in C} \frac{1}{\sqrt{AvgRank}} \quad (6)$$

where average facet rank AvgRank is calculated as

$$AvgRank = \frac{1}{|SFL|} \sum_{\text{l} \in \text{SFL}} Rank_{\text{e}} \quad (7)$$

where SFL is the set of all list in the facet group.

IV. EXPERIMENTAL ANALYSIS

4.1. Evaluation matrix

Precision recall facet: Precision recall facet (PRF): is used to combine the tree factors precision, recall and quality of facet terms and it is given as μ, $^{\omega}$ are used to adjust the emphasis among the above factors and F is system generated facet, F* is corresponding human generated facet.

$$PRF_{\alpha,\,\beta}(F, F^*) = \frac{(\mu^2 + \omega^2 + 1)prf}{\mu^2 rf + \omega^2 pf + pr} \quad (8)$$

Normalized discounted cumulative gain(n-DCG): Discounted Cumulative Gain(DCG) is the measure of usefulness and ranking quality of a search process. It is given as

$$DCG_{\text{p}} = \sum_{\text{i=1 to n}} \frac{2^{\text{rel}_\text{i}} - 1}{log(i+1)} \quad (9)$$

where reli is the graded relevance of the result at position i. Inorder to normalize DCG values an ideal ordering of the given query is needed. It is a monotonically decreasing function and is measured as the ratio of DCGi to Ideal Discounted Cumulative Gain (IDCGi) called n-DCG.

Purity aware n-DCG(fp-nDCG): In fp-nDCG measure, for each ground truth class ci' credit the first facet that is assigned to it. Here purity of each facet is further considered, and its weight is calculated as

Recall aware n-DCG(rp-nDCG): rp-nDCG calculation is based on all query facets. It's eight is calculated as

$$weight_{\mathrm{i}} = \frac{|c_{\mathrm{i}}' \cap c_{\mathrm{i}}|}{|c_{\mathrm{i}}|} * \frac{|c_{\mathrm{i}}' \cap c_{\mathrm{i}}|}{|c_{\mathrm{i}}'|} \quad (11)$$

Here $|c_{\mathrm{i}}' \cap c_{\mathrm{i}}| / |c_{\mathrm{i}}|$ is the percentage of item in ground truth ci' is matched to the current output facet ci.

Weighted precision recall facet(wPRF): wPRF is modified PRF in which it further accounts the different ratings associated with query facets, and it uses weighted facet term precision, recall, and clustering.

4.2. Dataset

Two datasets named UserQ and RandQ is used for experimental analysis over proposed framework. UserQ data set is a collection of queries about some user selected topics on real world data. Such 89 queries are considered and corresponding facets are identied, gave importance to facet lists more than texts. UserQ data set is prepared from the search log of commercial search engine. 105 English queries are considered for this.

4.3. Effect of list weighing method

According to subsection 3.2 the importance of a facet list is measured in terms of document matching weight and inverted document frequency. A list which contains a term frequently in it would show a high document weight for it, but at the same time some valid facets may degrade because of less frequent appearance. If including idfe in calculation this problem can solve. Document weight as obtained from (2) is the measure of facets which are on various lists and it improved the facet count in lists. Facet count is used to measure disjoint facet count in order to identify highly weighted facets. Figure 4. shows the effect of list weighting in facet mining in which various measurement methods are given in X-axis to various facet weights.

4.4. Effect of search result quantity

When the number of search results is increased then the facet groups are enriched with more lists and hence the quality of facets are improved Figure 5. Shows this. From section 3.2 it is clear that the facet count has an important role in the facet list weighing process. So if the results are more there is a possibility of facet count to be more. Specific properties whose facets match many products have high impurity

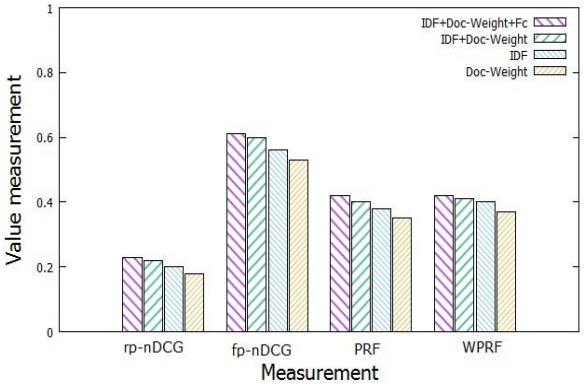

Fig. 4. Effect of various list weighting methods.

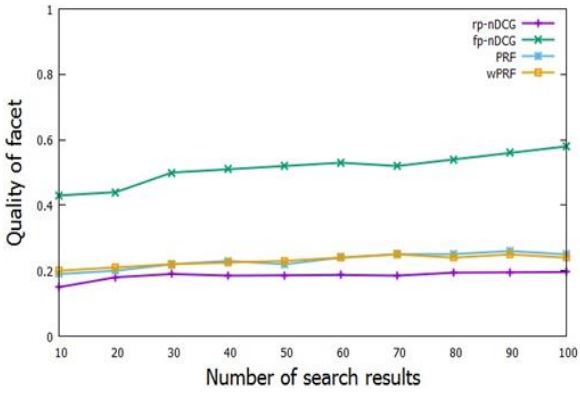

Fig. 5. Effect of search result quantity

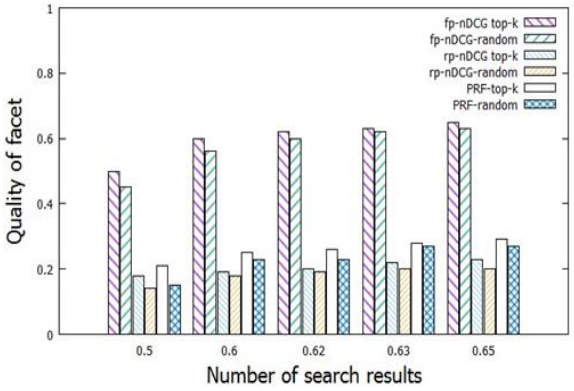

Fig. 6. Effect of search result quality.

4.5. Effect of search result quality

The quality of facets is increased with increased number of search results used by FMAD. If the number of search results increased the facet list count produced from them should be more. This in turn produces a variety of results and hence quality to facet. Figure 6. shows this.

V. CONCLUSION

The proposed Dynamic facet mining approach followed a series of phases such as facet clustering and grouping to drill down to the required items. Because of the abundance of facet it used a facet weighing and clustering scenario. Quality clusters are derived by considering disjoint facet count. Hamming distance is measured for cluster grouping by considering each facet group in a cluster. Average facet rank is calculated to list out top-k items which are specific on query and dispersed on query dimension. The Facet ordering which is used for dynamic dataset can be used to improvise qualitative and quantitative property calculations to feedback more confident segments.

REFERENCES

[1] J. Liu and D. Yan, Answering Approximate Queries over XML Data, IEEE Transactions on Fuzzy Systems, vol. 24, no. 2, pp. 288-305, 2016.

[2] E. J. Ruiz, V. Hristidis, and P. G. Ipeirotis, Facilitating Document Annotation using Content and Querying Value, IEEE transactions on knowledge and data engineering, vol. 26, no. 2, pp. 336-349, 2014.

[3] B. Kules and R. Capra, Creating Exploratory Tasks for a Faceted Search Interface, Procedure of HCIR, 2008, pp. 18-21, 2008.

[4] D. Vandic, S. Aanen, F. Frasincar, and U. Kaymak, Dynamic Facet Ordering for Faceted Product Search Engines, IEEE Transactions on Knowledge and Data Engineering, vol. 29, no. 5, pp. 1004-1016, 2017.

[5] Z. Dou, Z. Jiang, S. Hu, J.-R. Wen, and R. Song, Automatically Mining Facets for Queries from Their Search Results, IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 2, pp. 385-397, 2016.

[6] Q. Liu, E. Chen, H. Xiong, C. H. Ding, and J. Chen, Enhancing Collaborative Filtering by User Interest Expansion via Personalized Ranking, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 1, pp. 218-233, 2011.

[7] H.-J. Kim, Y. Zhu, W. Kim, and T. Sun, "Dynamic Faceted Navigation in Decision Making Using Semantic Web Technology, Decision Support Systems, vol. 61, pp. 59-68, 2014.

[8] Z. Wang, L. Shou, K. Chen, G. Chen, and S. Mehrotra, On Summarization and Timeline Generation for Evolutionary Tweet Streams, IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 5, pp. 1301-1315, 2015.

[9] M. A. Hearst, UIs for Faceted Navigation: Recent Advances and Remaining Open Problems, HCIR 2008: Proceedings of the Second Workshop on Human-Computer Interaction and Information Retrieval, pp. 13-17, 2008.

[10] J. C. Fagan, Usability Studies of Faceted Browsing: A Literature Review, Information Technology and Libraries, vol. 29, no. 2, pp. 58-66, 2010.

[11] M. Hearst, Design Recommendations for Hierarchical Faceted Search Interfaces, ACM SIGIR workshop on faceted search, pp. 1-5, 2006.

[12] D. Dash, J. Rao, N. Megiddo, A. Ailamaki, and G. Lohman, Dynamic faceted search for discovery-driven analysis, Proceedings of the 17th ACM conference on Information and knowledge management, pp. 3-12, 2008.

[13] D. Vandic, F. Frasincar, and U. Kaymak, Facet Selection Algorithms for Web Product Search, pp. 2327-2332, 2013.

[14] W. Kong and J. Allan, Extending Faceted Search to the General Web, Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pp. 839-848, 2014.

[15] O. Ben-Yitzhak, N. Golbandi, N. Har'El, R. Lempel, A. Neumann, S. OfekKoifman, D. Sheinwald, E. Shekita, B. Sznajder, and S. Yogev, Beyond Basic Faceted Search, pp. 33-44, 2008.

[16] C. Li, N. Yan, S. B. Roy, L. Lisham, and G. Das, Facetedpedia: Dynamic Generation of Query-Dependent Faceted Interfaces for Wikipedia, Proceedings of the 19th international conference on World wide web, pp. 651-660, 2010.

[17] Z. Yu, H. Wang, X. Lin, and M. Wang, Understanding Short Texts through Semantic Enrichment and Hashing, IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 2, pp. 566-579, 2016.

[18] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst, Faceted Metadata for Image Search and Browsing, pp. 401-408, 2003.

[19] S. Riezler, Y. Liu, and A. Vasserman, Translating Queries into Snippets for Improved Query Expansion, Proceedings of the 22nd International Conference on Computational Linguistics, pp. 737-744, 2008.
[20] S. Liberman and R. Lempel, Approximately Optimal Facet Value Selection, Science of Computer Programming, vol. 94, pp. 18-31, 2014.
[21] M. Jayapandian and H. Jagadish, Automated Creation of a Forms-based Database Query Interface, Proceedings of the VLDB Endowment, vol. 1, no. 1, pp. 695-709, 2008.
[22] S. R. Jeery, M. J. Franklin, and A. Y. Halevy, Pay-as-you-go User Feedback for Dataspace Systems, Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pp. 847-860, 2008.