SAODFT: Socially Aware Ontology Driven Approach for Query Facet Generation in Text Classification

Rituraj Ojha[1(✉)] and Gerard Deepak[2]

[1] Department of Metallurgical and Materials Engineering, National Institute of Technology, Tiruchirappalli, India
rituraj0480@gmail.com

[2] Department of Computer Science and Engineering, National Institute of Technology, Tiruchirappalli, India

Abstract. Query facet is a collection of topics or information that provides an overall summary of the user query. Only a few words or phrases from the relevant query facet can describe the user query better than the long sentences and this can be helpful during searches made on public search engines. In this chapter, a socially aware and ontology driven query facet generation framework is proposed. The proposed framework takes user query as input which is preprocessed and the obtained individual query words are integrated with domain ontologies. Furthermore, terms and events are integrated from the Twitter API which makes the model socially aware and relatable to real-world social events. The metadata is generated using OpenCalais and RDF Distiller, and is classified using Random Forest, taking individual query words as input and only the top 20% of the classified instances are taken. Similarly, the RandQ dataset is classified by taking individual query words as input using Random Forest. Finally, SemantoSim measure and Gamma Diversity measure is computed between the classified instances from the RandQ dataset and the instances coming from top 20% of the classified metadata and the entities passing through threshold value is formulated into a facet set by grouping them on decreasing value of the semantic similarity. The proposed approach achieves the highest average accuracy with the Precision of 90.61%, Recall of 93.18%, Accuracy of 91.89%, Purity of 0.95, F-Measure of 91.88%, nDCG value of 0.90, and low False Discovery Rate of 0.094 making it better approach than the baseline models.

Keywords: Facet recommendation · Metadata generation · Ontology driven approach · Query facet generation · Text classification

1 Introduction

The world wide web is populated with a large amount of information, including texts, images, videos, documents, etc. which keeps increasing every day. There is a large amount of data that a user can search on public search engines through his query. With increase in the amount of information available on the internet, it is getting harder and

S. Mekhilef et al. (Eds.): ICEEE 2022, LNEE 894, pp. 154–163, 2022.
https://doi.org/10.1007/978-981-19-1677-9_14

harder for users to find the relevant information pertaining to their submitted queries. In such cases, query facets play an important role helping the user to get relevant and most important information from the world wide web. Query Facets are collections of topics that summarize and describe an important and essential feature of the query. Facets can be phrases or a word. Several facets can be available for a query which describes the query from all the different perspectives. Hence, query facets can help to improve search results through its knowledge. For example, facets for query watches can describe the query in several distinct ways such as categories, styles, brands, gender and colors. Users can get important aspects of watches, such as brands like Rolex, Omega, etc. and also discover new information without wasting time by browsing through a lot of websites to get appropriate information. Users can also visit the women's section to gift a watch to their female friend, as recommended by the facets. The query facets can also help in searches for ambiguous queries like, Apple, where searches can contain results from fruit apple and also from technology brand Apple. For this, different query facets will be available to choose for both, which contains detailed information about fruit in one and technology brand Apple, in the other. Facets can also improve the diversity of the information by re-ranking the search results and eliminating the websites containing duplicate information. Hence, there is a need for a query facet generating tool which can be ontology driven and socially aware to work better than the already existing approaches. In the previous work, Ramya et al., has proposed an approach for the extraction of query facets by computing cosine similarity and using a high quality clustering algorithm [1].

Motivation: Query facets can summarize the entire user query from all the diverse perspectives which can help in getting better search results in the first page of the search engine. Hence, a better query facet recommendation framework is needed. Also, there are very few systems in this domain that are socially aware of the current real-world events and also return results with high accuracy. Therefore, building this system was a major motivating factor.

Contribution: The socially aware and ontology driven query facet generation approach is proposed. The framework takes the user query as input which is preprocessed and the domain ontologies are integrated with the obtained individual query words. Furthermore, terms and events are integrated from the Twitter API. The metadata is generated and is classified using Random Forest, taking individual query words as input and only the top 20% of the classified metadata instances are taken. Similarly, the RandQ dataset is classified by taking individual query words as input using Random Forest. Finally, SemantoSim measure and Gamma Diversity measure is computed between the classified instances from the RandQ dataset and the instances coming from top 20% of the classified metadata and the entities passing through the threshold value is formulated into a facet set by grouping them on decreasing value of the semantic similarity. The values of metrics namely, Precision, Accuracy, Purity, Recall, nDCG, and F-Measure are increased.

Organization: The flow of the remaining chapter is as follows. A condensed summary of the related works is provided in Sect. 2. Section 3 depicts the architecture of the proposed system. Section 4 describes the architecture implementation. Section 5 presents the evaluation of results and performance. The conclusion of the chapter is presented in Sect. 6.

2 Related Works

Kong and Allan [3] have proposed a methodology for query facet extraction using the empirical utility maximization approach. This approach increases the performance and is a precision-focused approach. The proposed approach only shows query facets for a few queries from all the input queries by calculating extraction performance for each query and giving only a few facets based on the extraction performance value. Jiang et al. [4] have proposed a technique for query facet mining using Knowledge Bases instead of using traditional methods of taking topics from top search results. Their approach mines initial facets from search engines and then entities from the Freebase knowledge base are integrated into it.

Chakraborty et al. [5] have proposed a model for faceted search of scientific articles on the internet. Their framework, FeRoSA, is based on the random walk model and it also organizes the scientific articles into facets based on their category. Ramya et al. [6] have proposed a technique for mining the query facets for user query automatically. The grouping of the facets happens by using a high quality clustering algorithm and calculating cosine similarity and the top items are collected as facets and recommended to the user.

Hu et al. [7] have proposed a method for diversifying the search results of the different users searching with the same query. The framework takes the help of query facets for the input user query and generates subtopics for each topic present in the facet. Vaishakhi and Regi [8] have proposed an approach for mining query facets with high quality by introducing their framework, namely, Anchor.

Radhakrishnan and Madhav [9] have proposed a query facet search engine for improving the search results of user queries. The engine will itself fetch the relevant facets for the search query and facets will be selected priority-wise based on their importance. Siddiqui et al. [10] have proposed a model to categorize the large documents using facets. This is done by extracting facets from these documents using their proposed framework, namely FacetGist which constructs a graph-based heterogeneous network to collect data.

3 Proposed System Architecture

The architecture for query expansion is depicted in Fig. 1. The proposed approach takes place in several steps. Initially, the User Query is taken as an input and is preprocessed using Tokenization, Lemmatization, stop word removal, and Named Entity Recognition (NER). Tokenization is the process of breaking the texts into small pieces called tokens. Byte Pair Encoding (BPE) is used for the tokenization process. Lemmatization involves grouping together several inflected kinds of the same word so that they are analyzed as one term. During stop word removal, the common ubiquitous words are removed as they add no value for the analysis and only increase the dimension of the feature set. NLTK python library is used for stop word removal. NER is the process of finding and categorizing the data or entity into predefined categories. After the preprocessing stage, individual query words are obtained.

These individual query words are integrated with domain ontologies. These domain ontologies are both static and dynamic. Static domain ontologies are ready-made ontologies which are already made using available domain ontologies to which the query as well as the dataset belongs to is used and the dynamic ontology is generated using user blogs, community forums, and contents from world wide web, using Onto Collab as a tool [11]. The next step involves integration of terms and events from the Twitter API. This makes the proposed methodology socially aware and relatable to real-world social events, as well as socially popular tags. Metadata is generated using OpenCalais and RDF Distiller. The metadata is classified using Random Forest, taking individual query words as training input. Only, top 20% of the classified metadata instances are selected. The reason for selecting only the top 20% is because the metadata obtained is extensively large and we only want the most relevant and consolidated entities from the metadata.

Random Forest is an ensemble machine learning model which can perform classification as well as regression, and any other tasks which can be done by constructing and training a very large number of decision trees. During the classification of metadata using Random Forest in the proposed methodology, the output class is the class which is selected by most decision trees. Random Forest is based on Bagging algorithm, where for each decision tree the model selects random bootstrap samples and a random subset of features from the training set.

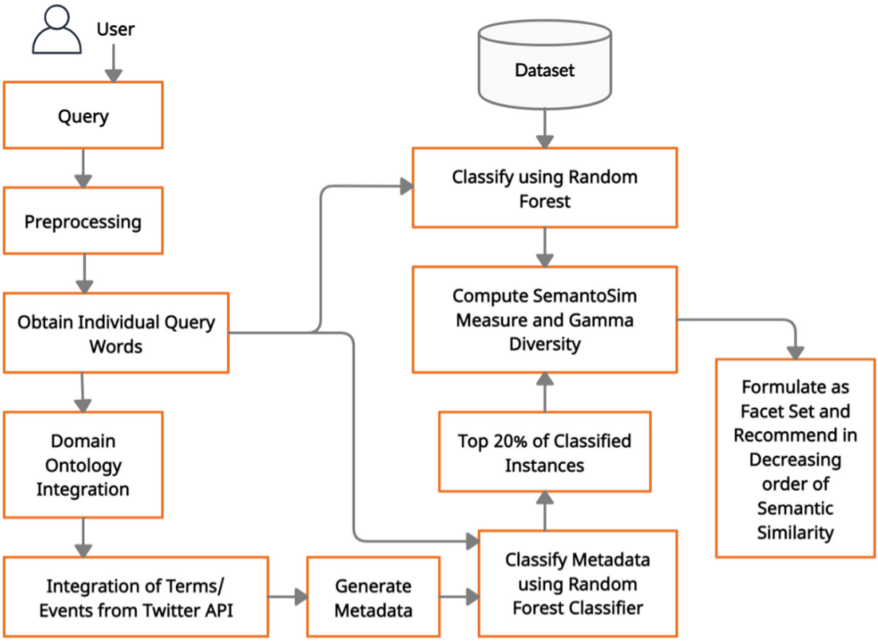

Fig. 1. Proposed system architecture

The dataset selected for this approach is RandQ. This dataset is built by Dou et al. [2] by collecting queries and building the query facets by extracting the top results from

the search engine using their tool, namely, QDMiner. The dataset is in XML format and 105 queries are taken from the dataset. Each query contains 100 documents and each document contains Document id, Title, description, URL, rank of the document, document text, and HTML list.

The dataset is classified by taking individual query words as input using Random Forest. Now, we will compute the SemantoSim measure and Gamma Diversity between classified instances from the dataset and the instances coming from the top 20% of the classified metadata. The threshold for the SemantoSim measure is 0.75 and the threshold for Gamma Diversity is 0.5. The entities passing through the threshold values are formulated into a facet set by grouping them in order of decreasing value of the semantic similarity, such that the higher value and most similar entities are kept first as compared to low value semantic similarity. To compute the semantic similarity and Gamma Diversity, a software agent is designed with the state and behaviour. The state will compute both the SemantoSim measure and Gamma Diversity, and behaviour will take the intersection of terms passed from semantic similarity and Gamma Diversity and group them as facets in decreasing value of semantic similarity which is then recommended to the user.

Gamma diversity as represented by Eq. (1), is the diversity of total species in a landscape, an island, or an area. The Alpha Diversity, or α in Eq. (1) represents the mean species diversity in the habitats of a local region. The Beta Diversity, or β in Eq. (1) represents the species diversity between two adjacent ecosystems.

$$\gamma = \alpha + \beta \tag{1}$$

SemantoSim is a semantic similarity measure proposed by Church & Hanks. Equation (2) represents the formula for computing SemantoSim Measure.

$$SemantoSim(x, y) = \frac{p(x, y)log\big[p(x, y)\big] + pmi(x, y)}{log\big[p(y, x)\big] + [p(x) * p(y)\big]} \tag{2}$$

4 Implementation

The dataset used in this proposed work is the RandQ dataset. It contains 105 queries in XML format. Every query contains 100 documents and each document contains Document id, Title, description, URL, rank of the document, document text, and HTML list. The implementation is done in the Google Collaboratory environment. The computer is equipped with 16 GB RAM and an i7 processor. Static domain ontologies are ready-made ontologies which are already made using available domain ontologies to which the query as well as the dataset belongs to is used and the dynamic ontology is generated using user blogs, community forums, and contents from world wide web, using Onto Collab as a tool.

Algorithm 1: Proposed SemUserProfiling Algorithm

Input: **Output:**	User Query and RandQ Dataset. Recommended Facet set.
Begin	
Step 1:	Query Q is subjected to query pre-processing. Q is tokenized, lemmatized, and NER and stop word removal is performed to obtain the individual query words Qw.
Step 2:	Qw is integrated with domain ontologies.
Step 3:	For each Qw as label: Qw ← Terms & Events from Twitter API End For.
Step 4:	Qwm ← Metadata generated for Qw terms using OpenCalais and RDF Distiller.
Step 5:	5.1: Model = RandomForestClassifier.fit(Individual Query Words Qw) 5.2: Model.transform(Metadata Qwm) 5.3: Qt = Select Top 20% of the classified Instances.
Step 6:	6.1: Model = RandomForestClassifier.fit(Individual Query Words Qw) 6.2: Model.transform(RandQ Dataset) 6.3: Qd = Classified Dataset Instances
Step 7:	For each a in Qt and b in Qd: Qres ← SemantoSim(a,b) > Thr. & GammaDiversity(a,b) > Thr.
Step 8:	Recommend Facet Set by grouping them based on decreasing value of the semantic similarity.
End	

5 Results and Performance Evaluation

The Performance of the proposed SAODFT (Socially Aware Ontology Driven Approach for Query Facet Generation in Text Classification) approach is measured by considering metrics namely Precision, Recall, Accuracy, and Purity. Other metrics including Normalized Discounted Cumulative Gain (nDCG), F-Measure and False Discovery Rate (FDR) are also calculated.

$$Precision = \frac{Retrieved \cap Relevant}{Retrieved} \tag{3}$$

$$Recall = \frac{Retrieved \cap Relevant}{Relevant} \tag{4}$$

$$Accuracy = \frac{Precision + Recall}{2} \tag{5}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$nDCG = \frac{DCG_{\propto}}{IDCG_{\propto}} \tag{7}$$

$$DCG = \sum_{i=1}^{\propto} \frac{rel_i}{log_2(i+1)} \tag{8}$$

$$False\ Discovery\ Rate = 1 - Positive\ Predictive\ Value \tag{9}$$

$$Purity = \frac{1}{N} \times \sum_{m \in M} max_{d \in D} |m \cap d| \tag{10}$$

The Eqs. (3), (4), and (5) represent Precision, Recall and Accuracy, respectively. Furthermore, the Eqs. (6), (7), and (8) represent F-Measure, nDCG and Discounted cumulative gain, respectively. Equations (9) and (10) represent FDR and Purity, with 'M' clusters and 'D' classes partitioning N number of datapoints. The reason for considering Accuracy, Precision, Recall, Purity and F-Measure metrics for evaluation is because they measure the relevance of the results and FDR computes the number of false discoveries made by the proposed system. The nDCG represents the diversity of the relevant results.

Table 1. Comparison of performance of the proposed SAODFT with other approaches for the RandQ dataset.

Search Technique	Average Precision %	Average Recall %	Accuracy %	Purity	nDCG	F-Measure	FDR
QDMiner [2]	70.14	74.14	72.14	0.88	0.68	72.08	0.298
AEFTM [1]	71.14	77.18	74.16	0.92	0.72	74.04	0.288
Proposed SAODFT	90.61	93.18	91.89	0.95	0.90	91.88	0.094
Eliminating Metadata Generation from the Proposed SAODFT	81.15	85.12	83.14	0.86	0.81	83.08	0.188
Eliminating SemantoSim Measure from SAODFT	83.44	86.17	84.81	0.84	0.85	84.78	0.165
Eliminating Gamma Diversity Index from SAODFT	82.17	85.78	83.98	0.83	0.86	83.93	0.178

Table 1 represents the performance comparison of the proposed SAODFT with that of base paper along with other approaches considered by using elimination techniques. It is

evident from the table that the proposed approach achieves the highest average accuracy with the Precision of 90.61%, Recall of 93.18%, Accuracy of 91.89%, Purity of 0.95, F-Measure of 91.88%, nDCG value of 0.90, and low FDR of 0.094. The baseline model for this paper AEFTM [1], achieves much less average accuracy, getting Precision of 71.14%, Recall of 77.18%, Accuracy of 74.16%, Purity of 0.92, F-Measure of 74.04%, nDCG value of 0.72, and FDR of 0.288. The performance of the SAODFT framework is better because of several reasons. Firstly, the metadata generation increases the real-world knowledge. Secondly, domain ontology integration enhances the probability of incidence of concepts which are highly relevant to the base framework. Thirdly, the Twitter API makes the framework socially aware of real-world events. Lastly, SemantoSim measure ensures that the relevance of the topics is high and Gamma Diversity ensures that the diversity of the relevant topics is also high. Eliminating any of the used methodologies in the proposed SAODFT framework will lead to a significant decrease in the average accuracy of the framework.

In QDMiner [2], the approach extracts the query facets by gathering the similar topics from the top search results on the search engine. This can lead to getting irrelevant noise and duplicate results. The AEFTM [1] is not as good as the proposed framework because it uses only the text mining approach. Also, they are using only Clustering algorithm with cosine similarity which is highly naïve and traditional. They are using domains from the dataset, but the proposed framework uses domains from dataset along with the domain from social network, Twitter. Hence, the proposed SAODFT framework works much better than the baseline models.

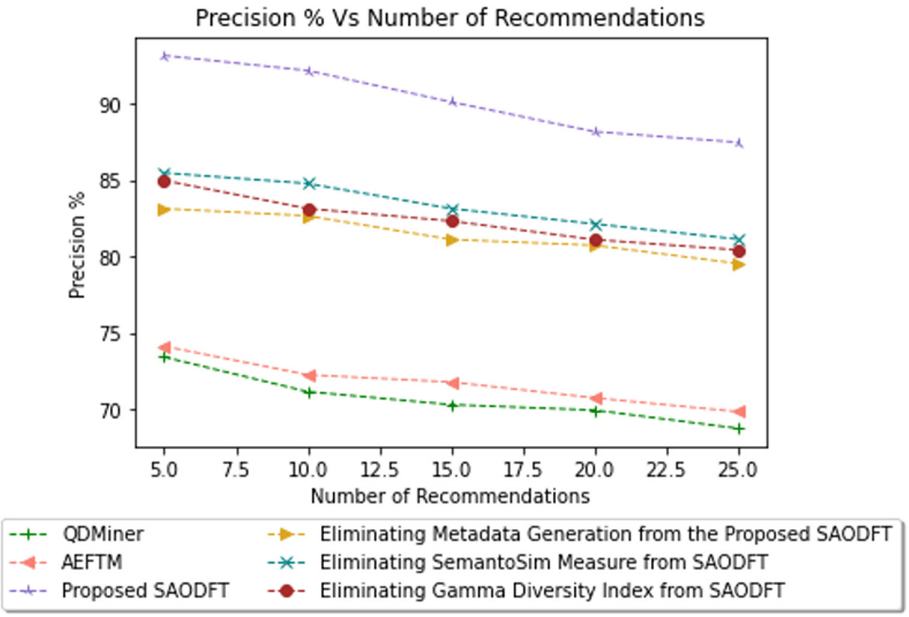

Fig. 2. Precision % vs number of recommendations

Figure 2 represents Precision % vs number of recommendations line graph for all the approaches. It is evident from the figure that the proposed SAODFT framework achieves the highest Precision % for any number of recommendations. Even after the elimination of one methodology used in the proposed approach, the precision remains greater than the two baseline models, AEFTM [1] and QDMiner [2]. Hence, we can conclude that the proposed model is better and more accurate than the baseline models.

6 Conclusions

In this chapter, a socially aware ontology driven approach for query facet generation in text classification is proposed. The increase in the amount of information available on the internet has made it harder for users to find the relevant information pertaining to their submitted queries. Therefore, query facets play an essential role in helping the user to get relevant and most important information from the world wide web. Query Facets are collections of topics that summarize and describe an important and essential feature of the query. The proposed framework takes user query as input which is preprocessed and the obtained individual query words are integrated with domain ontologies. Furthermore, terms and events are integrated from the Twitter API. The metadata is generated using OpenCalais and RDF Distiller, and is classified using Random Forest, taking individual query words as input and only the top 20% of the classified instances are taken. Similarly, the RandQ dataset is classified by taking individual query words as input using Random Forest. Finally, SemantoSim measure and Gamma Diversity measure is computed between the classified instances from the RandQ dataset and the instances coming from top 20% of the classified metadata and the entities passing through threshold value is formulated into a facet set by grouping them on decreasing value of the semantic similarity. The proposed approach achieves the highest average accuracy with the Precision of 90.61%, Recall of 93.18%, Accuracy of 91.89%, Purity of 0.95, F-Measure of 91.88%, nDCG value of 0.90, and FDR of 0.094. The overall accuracy of the proposed technique is much better than the existing approaches.

References

1. Ramya, R.S., Raju, N., Pushpa, C.N., Venugopal, K.R., Iyengar, S.S., Patnaik, L.M.: Automatic extraction of facets for user query in text mining [AEFTM]. Int. J. Emerg. Technol. **11**(2), 342–350 (2020)
2. Dou, Z., Jiang, Z., Hu, S., Wen, J.R., Song, R.: Automatically mining facets for queries from their search results. IEEE Trans. Knowl. Data Eng. **28**(2), 385–397 (2015)
3. Kong, W., Allan, J.: Precision-oriented query facet extraction. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 1433–1442 (2016)
4. Jiang, Z., Dou, Z., Wen, J.R.: Generating query facets using knowledge bases. IEEE Trans. Knowl. Data Eng. **29**(2), 315–329 (2016)
5. Chakraborty, T., Krishna, A., Singh, M., Ganguly, N., Goyal, P., Mukherjee, A.: Ferosa: a faceted recommendation system for scientific articles. In: Bailey, J., Khan, L., Washio, T., Dobbie, G., Huang, J.Z., Wang, R. (eds.) Advances in Knowledge Discovery and Data Mining: 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016,

Proceedings, Part II, pp. 528–541. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-31750-2_42
6. Bedi, P., Goyal, S.B., Rajawat, A.S., Shaw, R.N., Ghosh, A.: A framework for personalizing atypical web search sessions with concept-based user profiles using selective machine learning techniques. In: Bianchini, M., Piuri, V., Das, S., Shaw, R.N. (eds.) Advanced Computing and Intelligent Technologies. LNNS, vol. 218, pp. 279–291. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2164-2_23
7. Hu, S., Dou, Z.C., Wang, X.J., Wen, J.R.: Search result diversification based on query facets. J. Comput. Sci. Technol. **30**(4), 888–901 (2015)
8. Kumar, A., Das, S., Tyagi, V., Shaw, R.N., Ghosh, A.: Analysis of classifier algorithms to detect anti-money laundering. In: Bansal, J.C., Paprzycki, M., Bianchini, M., Das, S. (eds.) Computationally Intelligent Systems and their Applications. SCI, vol. 950, pp. 143–152. Springer, Singapore (2021). https://doi.org/10.1007/978-981-16-0407-2_11
9. Radhakrishnan, A., Madhav, M.L.: Query facet engine for easier search results. In: 2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1–5 (2017)
10. Siddiqui, T., Ren, X., Parameswaran, A., Han, J.: Facetgist: collective extraction of document facets in large technical corpora. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 871–880 (2016)
11. Pushpa, C.N., Deepak, G., Thriveni, J., Venugopal, K.R.: Onto Collab: strategic review oriented collaborative knowledge modeling using ontologies. In: 2015 Seventh International Conference on Advanced Computing (ICoAC), pp. 1–7 (2015)