Towards Human-in-the-Loop Based Query Rewriting for Exploring Datasets
Genoveva Vargas-Solar, Mehrdad Farokhnejad, Javier A Espinosa-Oviedo

To cite this version:
Genoveva Vargas-Solar, Mehrdad Farokhnejad, Javier A Espinosa-Oviedo. Towards Human-in-the-Loop Based Query Rewriting for Exploring Datasets. Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference, Mar 2021, Nicosia, Cyprus. hal-03183517

HAL Id: hal-03183517
https://hal.science/hal-03183517
Submitted on 27 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Towards Human-in-the-Loop Based Query Rewriting for Exploring Datasets

Genoveva Vargas-Solar  
CNRS, LIRIS-LAFMIA  
Lyon, France  
genoveva.vargas-solar@liris.cnrs.fr

Mehrdad Farokhnejad  
Univ. Grenoble Alpes, Grenoble INP, CNRS, LIG  
Grenoble, France  
mehrdad.farokhnejad@univ-grenoble-alpes.fr

Javier A. Espinosa-Oviedo  
Univ. Lyon 2, EIRIC-LAFMIA  
Lyon, France  
javier.espinosa@acm.org

ABSTRACT
Data exploration promotes a new querying philosophy that gradually converges into queries that can be used to exploit raw data collections according to data explorers (i.e., users) expectations. Data exploration aims to guide the understanding of data collections with different rawness degree and define the type of questions that can be asked on top of them, often through interactive exploration processes. This paper introduces a human-guided data exploration approach defining exploration operations that result in different types of factual and analytic queries. Our first results include a proposal of query morphing and queries as answers strategies. This paper describes an experiment setting used for testing the data exploration techniques.

1 INTRODUCTION
The evolution in querying, information retrieval and human-computer interaction has led to the shift of interest from traditional Query-Response paradigm to actual human intelligence systems. Approaches such as interactive query expansion (IQE) [3, 9, 19] have shown the importance of data consumers in the data exploration process. Users’ intention helps to navigate through the unknown data, formulate queries and find the desired information. In most of the occurrences, user feedback acts as vital relevance criteria for next query search iteration. Such novel requirements of modern exploration driven processes call to rethink data querying processes.

Traditional data management systems assume that when users ask a query (i) they have good knowledge of the schema, meaning and contents of the database, and (ii) they are sure that this particular query is the one they wanted to ask. In short, it is assumed that users know what they are looking for. In response, systems like DBMS, always try to produce correct and complete results [7]. These assumptions are becoming less true as the volume and diversity of data grow, and as raw datasets representing phenomena observations, rather than facts, need to be explored by data scientists and other users. First, the structure and content of the database are hard to understand. Second, finding the right question to ask is a long-running and complex task, often requiring a great deal of experimentation with queries, backtracking query results, and revision of results at various points in the process [18]. Existing systems have limited provisions to help the users to reformulate their queries as they evolve with the search progression [10].

This paper proposes a data exploration approach that defines a loop where, given a user intention expressed using terms and raw data collections, the exploration strategies propose different types of possible queries that can be asked on top of data, and that potentially correspond to user’s expectations; interacts with the user for refining intentions based on the proposed queries and starts the loop again until the queries proposal converges with user expectations.

Different data exploration pipelines can be defined combining different exploration techniques for performing specific data exploration tasks. The rest of the paper is organised as follows. Section 2 summarises the related work done in the field of data exploration proposing a classification of approaches and techniques. Section 3 provides a detailed description of the approach. Section 4 describes our experimental setting including the dataset and its pre-processing steps. Section 5 concludes the paper and discusses future work.

2 RELATED WORK
Data exploration calls for combining different exploration, querying and processing methods and strategies proposed in diverse domains. Therefore we performed a systematic review to classify them (see Figure 1). According to a systematic review we performed, we propose a classification of existing data exploration techniques and methods. The classification consists of facets representing an aspect of data exploration and dimensions that denote the concepts that define each facet. As shown in Figure 1, the facets classify: (F1) the type of queries addressed by existing work; (F2) the type of algorithms used for exploring data collections; (F3) the knowledge domain of data collections and data types; (F4) the exploration processes done with human intervention; and (F5) data exploration techniques and systems conceived for understanding raw datasets content.

Since exploration can put different types of queries in action, facet F1 classifies the types of queries that are defined and used in different works that exploit datasets. The spectrum goes from “classic” keyword and relational queries evaluated on top of more or less curated datasets, to data processing operations on raw datasets (e.g., descriptive statistics). In this spectrum, these types represent families of queries that can include aggregation, clustering operations. We mainly identify “query by example” techniques useful particularly in cases where the knowledge about the datasets’ content is too weak (see d1.8). Query by example is an intuitive way to explore data, so many techniques are applying it to data exploration. Examples can either represent approaches like reverse engineering querying and queries like query morphing or queries. We also note that data exploration is a loop that obtains approximated results and the techniques
are specialised according to the type of data models (relational, graph, semi-structured, text, multimedia).

Depending on the domain, works propose algorithms rather than operators (like in relational contexts) to process datasets and to discover and derive a precise statistical understanding of their content (facet F2). Algorithms sometimes depend on the type of data structures used for representing data. For example, there are algorithms for processing graphs (centrality, pathfinding, etc.) or querying tables (selection, projection, etc.). Many works use well-known heuristics, data mining, machine learning, artificial intelligence algorithms for processing datasets, and insight into their content. Finally, other works propose their strategies without adhering to a specific domain.

The vision of data exploration in this work is that it should be a human-guided process. Therefore, we have studied techniques where humans intervene to adjust and guide the process of receiving information (d.4.5). We studied works on group recommendation, consensus functions, group preference and group disagreement. These study address objectives like designing consensus functions that aggregate individual group members’ preferences to reflect the overall group’s preference for each item [1, 4, 13] or disagreement about an item [16]. Consensus functions can be applied within a data exploration process given where a user can agree and disagree about the proposed queries; the system can recommend queries according to given constraints that can be interpreted as preferences.

According to our classification, facet F5 considers dimensions that represent exploration techniques. Regarding exploration query expression (d.5.1), we have identified three types of approaches: multi-scale query processing for gradual exploration; query morphing to adjust for proximity results; queries as answers as query alternatives to cope with lack of providence. Results filtering (d.5.2) addresses analysis and visualisation to give insight to data content. Finally, data exploration systems & environments (d.5.3) are tailored for exploring data incrementally and adaptively.

Concerning data exploration techniques, M. L. Kersten et al. [13] have compiled five methods to explore data sets querying: one-minute DB kernels, multi-scale queries, result set post-processing and query morphing and queries as answers. These methods revisit fundamental characteristics of existing systems like the notion of results completeness and correctness promoted by traditional databases, splitting queries execution on different fragments of a database, precision of queries, and one-shot computation of query results. These query systems provide a broader (i.e., less precise but with a broader scope) approach, discarding exactness and completeness for speed and a more global vision of the data.

Finally, facet F3 classifies the type of datasets used to test different exploration techniques and approaches. Datasets content is often textual and with different rawness degree (newspapers, micro-texts from social networks) and already processed content using NLP (Natural Language Processing) techniques and represented as graphs or tables. Other datasets are built by collecting observations monitored using, for example, IoT infrastructures. These data sets contain records of measures or even video or images.

We have the following remarks about what we have studied in state of the art. Data exploration pipelines are mostly ad-hoc, implemented in an artisanal manner, and partially human-guided. Machine learning, analytics and querying techniques (e.g., query by example, queries as answers, etc.) are complementary. We observed that no existing system integrates them so that data scientists can develop exploration pipelines that can thoroughly understand data and its analytics potential. Therefore, there is room for proposing approaches for each of them, defining rules on how they can be combined within data exploration pipelines and integrating them to provide a whole data exploration environment.

3 QUERYING PIPELINES FOR EXPLORING DATASETS

Figure 2 shows our general approach based on query rewriting techniques and summarised as follows: “given an initial query, provide sets of queries that can help data consumers better exploit data collections”. The approach considers that data collections are textual and indexed (not necessarily cleaned) and the representative vocabulary used in their content has been extracted and classified. For example in a crisis management scenario, the classes are events (e.g., someone looks for shelter, a building has been damaged) and actions (e.g., a hotel provides shelter for victims, people is approaching a damaged building to search victims).

The approach is intended to rewrite initial keyword queries by morphing expressions to produce results that can retrieve representative insights into these collections’ content. The rewriting process is gradual and interactive, where the user expresses an initial expression, and the exploration process provides new queries associated with content samples that can give insight into the content of the dataset. The alternative queries are assessed and adjusted by the user. Then, the exploration process is triggered again until a set of queries is chosen to be evaluated to produce results. Results produced by different exploration strategies can also be used as input to others. For example, query morphing’s output can be used as an input for the queries as answers.

The next sections describe two rewriting techniques query morphing and queries as answers (expansion) that we have proposed for exploring datasets.

3.1 Query morphing

Query morphing is the process of rewriting conjunctive and disjunctive keyword queries, by adding terms, to increase the possibility of exploring the most number of items in a collection. We proposed and implemented a “query morphing” pipeline that can help the data scientist better precise her query (see Figure 3). Our query morphing pipeline uses a vocabulary and Wordnet to look for associated terms and synonyms that help expand the terms to enhance the chance of matching it with relevant data items in the target collection. The pipeline is described as follows. Given a conjunctive and disjunctive keyword query represented as an expression tree go through the tree in depth-first until finding a leaf representing a term and then:

1. Use a vocabulary representing the dataset content and Wordnet seeking for:
   (a) equivalent terms and generate a node with the operator and then connect the initial term with the equivalent terms in a conjunctive expression subtree.
   (b) more general terms and connect the initial term with these terms in a disjunctive expression subtree.
   (c) assess and adjust the morphed query by the user and eventually restart the expansion process. The assessment process includes pondering the terms and exploring result samples to see potential results.
For example, for a query "victims AND missing AND shelter", using Wordnet¹, the query can be expanded as follows: "(victim OR casualty OR "unfortunate person") AND (missing OR absent) AND (shelter OR protection OR housing)". The key is using a concept ontology or glossary and can find the maximum equivalent and general terms. In this example, we only found equivalent terms. The user can then mark which terms should or should not be included in the expanded query. She can also test different combinations of the query and compare the results to see which morphed query can produce the results that best respond to her expectations.

Once the new query expression has been rewritten, as done in information retrieval techniques, we use the inverted index to find the corresponding documents where the query terms occur. Then, we use the frequency matrix to compute the final result set tagged with precision and recall measures.

³ Query as answers

Given an initial conjunctive/disjunctive keyword query, the query is rewritten and transformed into several queries by extending it with general and more specific terms, synonyms, etc., and by exploiting the knowledge domain (see Figure 3). The result is a set of possible alternative queries with associated sample results so that the user can choose which ones to execute.

In our approach, the initial query is represented by an expression tree (intermediate representation) where nodes are conjunction and disjunction operators and leaves are terms. During the rewriting process, the tree is modified by adding new types of nodes and tagged arcs. New nodes represent "and" and "or" nodes that do not belong to the initial query and more general/specific terms associated with an "initial" term. These new nodes are connected with the nodes of the initial query by tagged arcs. A tag can indicate whether it connects a node with a conjunction or disjunction of more general/precise terms.

---

¹http://wordnetweb.princeton.edu/perl/
For example, consider an initial query linking three terms $t_1$, $t_2$ and $t_3$ with a conjunctive and disjunctive initial query (see Figure 3). It is then re-written in a new query represented by a tree that extends the query with terms that can be synonyms or related terms to $t_1$ and $t_2$. Three possible queries are derived: $Q_1^1$ which provides an alternative to $t_1$ with a new complex query with a synonym/or an associated term $t_4$, saying that we can look for "$t_1$ or $t_1$ and $t_4$". Similarly, $Q_1^2$ provides an alternative to $t_2$ saying that we can either look for $t_2$ or $t_2$ and $t_6$, which could be a synonym or a related term. Finally, $Q_1^3$ is a complex query that integrated $Q_1^1$, $Q_1^2$ with $t_3$ of the initial query.

The following steps are performed for computing queries alternatives where every step aims at deriving the initial query into queries that add knowledge. For each leaf in the expression tree of the query:

1. Use a vocabulary (extracted from the dataset content) and Wordnet seeking for:
   - equivalent terms and generate a node with the operator and then connect the initial term with the equivalent terms in a conjunctive expression subtree;
   - more general terms and connect the initial term with these terms in a disjunctive expression subtree.
2. Use a frequency matrix for looking for terms that are often associated with the initial term with a specific frequency and getting a sample of documents that can belong to query results.

The user can choose those queries that best target her expectations. A history of queries is maintained that can be reused for suggesting or pre-calculating morphing or query as answers results or for adjusting the chosen query set with new queries as the dataset evolves.

4 EXPERIMENTS

To experiment our general approach, let us consider a disaster management scenario where various data collections are produced during the life cycle of the disaster and must be explored to organise relief, resilience and duty of memory actions. The scenario we use is related to disaster management under a horizontal organisation ², where civilians take active action when an event happens (e.g., earthquake, flooding, fire) and continue to influence decision making during the other phases of its management. In this context, social media is a fast-paced channel used by affected people to describe their situation and observations, seek information, specify their requests, and offer their voluntary assistance: providing actionable information [17, 21]. Critical data is continuously posted on social media like Twitter, during the disaster life cycle (the event, relief, resilience, duty of memory).

During such life-threatening emergencies, affected and vulnerable people, humanitarian organisations, and other concerned authorities search for information useful to provide help and prevent a crisis. Nobody has control over the type of data exchanged by actors. These data are crucial in making critical decisions like saving lives, searching people, and providing shelter and medical assistance. For this reason, it is required to explore past and present in an agile manner to find hints to make decisions and act individually and collectively. Social network data collections can include reports on architectural and building environment damages and volunteers informing them that they have answered calls for help (see Figure 4).

The question is whether these data collections can help to find (i) causal correlations, for example, is it possible to know given a post asking for help whether actions have already been

²The phenomenon of organisation of civilians under horizontal and marginal groups has come up in different countries during rivers flooding, annual landslides in diverse regions particularly in Latin America.
Figure 4: Exploring crisis social network posts during crisis

4.1 Dataset preparation

Among social media studies, most of them focus on Twitter, mainly because of its timeliness and availability of information from a large user base. We use CrisisNLP [11] labelled and unlabelled datasets. The datasets contain approximately 5 million unlabelled and 50K labelled tweets. The size of this dataset is about 7 gigabytes. The datasets consist of various event types such as earthquakes, floods, typhoons, etc. The datasets were collected from the Twitter streaming API using different keywords and hashtags during the disaster. The tweets are labelled into various informative classes (e.g., urgent needs, donation offers, infrastructure damage, dead or injured people) and one not-related or irrelevant class. Table 5 shows a sample of some labelled tweets from data collection.

Data Preprocessing. Since the tweet texts are brief, informal, noisy, unstructured, and often contain misspellings and grammatical mistakes, preprocessing must be done before using them in further analysis. Moreover, due to Twitter’s 140 characters limit restriction, Twitter users intentionally shorten words using abbreviations, acronyms, slang, and sometimes words without spaces; hence we need to normalise those OOV terms. [11]. Presence of duplicates can result in an over-estimation of the performance of retrieval/extraction methodologies. Therefore, we eliminated duplicate tweets using ‘remove duplicates toolkit’ by Excel. Currently, we use 73562 unlabelled data set related to 2014. We performed the following preprocessing steps to clean the micro-documents:

- We removed stop words (e.g., ‘a’, ‘at’, ‘here’), non-ASCII characters, punctuations (e.g., ‘. ‘, ‘!’), URLs (e.g., ‘http://t.co/24Db32z04U’), hashtags (e.g., ‘#Napaquake!’) and Twitter reserved words (e.g., ‘RT’, ‘via’).
- We further tokenized the tweets using nltk.tokenize library [24].
- We performed stemming using the WordNet Lemmatizer library [24]: e.g. troubled (trouble).
We applied supervised techniques such as Random Forest [5], Support Vector Machines (SVM) and Convolutional Neural Network (CNN) [6, 8] to classify the tweets of our experiment dataset.

4.2 Testing query morphing

We implemented the "query morphing" process that we proposed to help the data scientist better precise her query, or define several queries representing what she is looking for. Our query morphing algorithm uses Wordnet to look for associated terms and synonyms that help expand the terms to enhance the chance of matching it with relevant tweets in the target collection.

For assessing expanded term quality, we have compared the results as the original paper [11, 15]. The results obtained by CNN model are better than traditional techniques, and we were able to obtain the same results as the original paper [11, 15] (see Table 1).

Table 1: Accuracy, Precision, Recall and f-score of CNN model with respect to California Earthquake and Nepal Earthquake crisis tweet data.

<table>
<thead>
<tr>
<th>Datasets/SYS</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>f-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>California Earthquake</td>
<td>92.72</td>
<td>96.53</td>
<td>90.00</td>
<td>88.23</td>
</tr>
<tr>
<td>Nepal Earthquake[12]</td>
<td>89.31</td>
<td>91.25</td>
<td>91.87</td>
<td>91.85</td>
</tr>
</tbody>
</table>

4.3 Expanding the query

We implemented the "query expansion" process that we proposed to expand the query by adding related terms from the expanded query terms obtained from the classified query. For assessing extended term quality, we have compared the results as the original paper [11, 15]. The results obtained by CNN model are better than traditional techniques, and we were able to obtain the same results as the original paper [11, 15].
In this section, we presented an ablation study about the performance of our proposed classification based query morphing method. We use the available crisisNLP pre-trained word embedding via word2vec method [11] to obtain query and expansion terms vectors. In the vector space model, all queries and terms are represented as vectors in dimensional space 300. Documents similarity is determined by computing the similarity of their content vector. To obtain a query vector, we represent keywords in user queries as vectors, and then sum all the keyword vectors followed by averaging them. For our analysis, we calculated the average similarity between the query vector (Q_vector) and ‘m’ keyword vectors obtained for a given query (T_vector) by using the formula of similarity Sim given in equation 1.

$$Sim(CTs, Query) = \frac{\sum_{i=1}^{m} \text{Cosine}(Q\_vector, T\_vector[i])}{m}$$  \hspace{1cm} (1)

where CTs are candidate terms, 'm' is a hyper-parameter in query expansion-based retrieval, which shows the number of expansion terms (ET), using as reference the studies[2, 22]. We set the number of expansion terms to 10, 20 and 30 (ET@10, ET@20, ET@30). We repeat this task for 100 queries and report the mean of average of each ET set in table 2. The experimental results show that the morphed query expanded with new terms obtained from the classified query morphing model are more similar and relevant than the non-classification model. The ET@10, ET@20 and ET@30 scores of our proposed classification model surpassed the transition non-classification based model. Also, we observe that when we set the number of expansion terms to 10, we achieve the best performance.

Currently, we used pseudo relevance feedback. This method automates the manual part of relevance feedback. It is assumed that the user takes top-m ranked morphed query terms returned by the initial query as relevant to expand her query. Results scoring must be completed with user feedback that finally guides the process. We have proposed a solution for exploring scientific papers through an experiment defining a set of exploration queries. Results were assessed by scientists of the National Institute of Genetic Engineering and Biotechnology, Tehran, Iran and Golestan University of Medical Sciences. Scientists provided feedback about exploration operations through questionnaires that are processed for obtaining satisfaction metrics. We are currently defining a crowd-based setting for obtaining feedback in the case of crisis datasets. The idea is to work with different groups of users (victims, volunteers, logistics decision-makers, police, medical staff) and queries to assess exploration results.

5 CONCLUSION AND FUTURE WORK

This paper introduced a general datasets exploration approach that includes human in the loop. The current approach includes two exploration techniques (i.e., query morphing and queries as answers) to help define queries that can fully explore and exploit a dataset. They are complementary query rewriting techniques where initially expanding a query can help adjust the terms used for exploring a dataset and then produce possible combinations of terms with possible queries that can lead to different scopes. In both cases, the user finally chooses a set of representative queries to her interests and the produced results that target her expectations. We have tested query morphing in the case of crisis dataset exploration, where people involved in a critical event either as victim or volunteers can define queries for retrieving information to look for or provide help.
Our future work includes modeling query exploration pipelines that can combine different techniques for exploring data collections. We will also propose ways of morphing and giving queries as answers where queries can be analytical or imply quantitative data views.

REFERENCES


