Computing query sets for better exploring raw data collections

Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, Javier-Alfonso Espinosa-Oviedo

Computing query sets for better exploring raw data collections

Genoveva Vargas-Solar, José-Luis Zechinelli-Martini, and Javier-Alfonso Espinosa-Oviedo

Abstract—This paper proposes an approach for helping data scientists express queries that can help them understand the content of raw data collections. The proposal is based on query rewriting techniques that given an initial query they can provide sets of queries that can help data scientists to better exploit data collections. We apply our approach in the context of the project *Exploring the History of Climate Change in Latin America through Newspaper Articles* funded by the Panamerican Institute of Geography and History (PAIGH) where data scientists explore newspapers to track articles reporting climatologic events.

Index Terms—Big data, query rewriting, ontologies.

1 Introduction

Technological advances have enabled to see the big picture about the Earth and its climate on a global scale by collecting many different types of information. Earth observation data collections will underpin the future earth program with a huge volume of various types of data and will play an important role in academia and decision making. Earth observation data has the 4V features (volume, variety, veracity, and velocity) of Big Data [1]. Data collections are exported under different releases with different sizes, formats (e.g., csv, text, excel), sometimes with various quality features. Tools helping to understand, consolidate and correlate data collections are crucial. We have collected and provided data collections that can help to reconstruct the history of climate change in Latin America.

The objective of data querying is to obtain all the data tuples respecting a defined often in the objective of answering a related question with correct and complete results. This means knowing what data is in the database and in what structure. In raw digital data collections this cannot be guaranteed. Often users are not sure which patterns they want to find and can be exploitable to answer their questions.

This paper proposes an approach for helping data scientists express queries that can help them understand the content of data collections. The proposal is based on query rewriting techniques that given an initial query they can provide sets of queries that can help data scientists to better exploit data collections. We apply our approach in the context of the project *Exploring the History of Climate Change in Latin America through Newspaper Articles* funded by the Panamerican Institute of Geography and History (PAIGH) where data scientists explore newspapers to track articles reporting climate events.

- *G. Vargas-Solar is Senior Scientist of the French Council of Scientific Research, LIG-LAFMIA Labs. France. E-mail: see http://www.vargas-solar.com*
- *J.L. Zechinelli Martini is associate professor of the Fundacion Universidad de las Améericas Puebla, in Mexico; and J.A. Espinosa-Oviedo is scientist at the LAFMIA lab.*

Manuscript received March 1, 2016.

The remainder of the paper is organised as follows. Section 2 introduces the general approach and context of the work regarding the exploration and analysis of newspaper articles. Section 3 describes our strategy for rewriting queries in order to derive new ones that can better target results using domain knowledge and knowledge about the content of the newspapers digital collections. Section 4 discusses related work concerning query rewriting techniques. Finally, Section 5 concludes the paper and discusses future work.

2 Exploring and analyzing newspaper articles

The work presented in this paper is part of a bigger project financed by the Panamerican Institute of Geography and History (PAIHG). The general principle of the problem addressed by the project has three main aspects as shown in figure 1.

First it introduces a data collection challenge that we characterized with the metaphor "newspaper archaeology". We have worked with the national libraries of Mexico, Colombia, Ecuador and Uruguay to access to their newspapers digital collections to discover articles talking about climatologic events happened between the XVIII and the XIX centuries. The problem also addressed the construction of a vocabulary used on those articles identified as reporting a meteorological event. Note that we wanted to discover linguistic variations in different Latin American countries used for describing climate events. The use of language and its variations can give a picture of the perception of civilians about these events, their consequences and associated explanations.

The second aspect of the problem concerns the processing of the articles potentially identified within newspapers to validate that they really report climatologic events. When this is the case, articles are semi-automatically tagged with meta-data that can provide an abstract representation of their content specifying what type of event they describe, where and when did the event happen? How long did it last? and which was their geographical scope / extension?

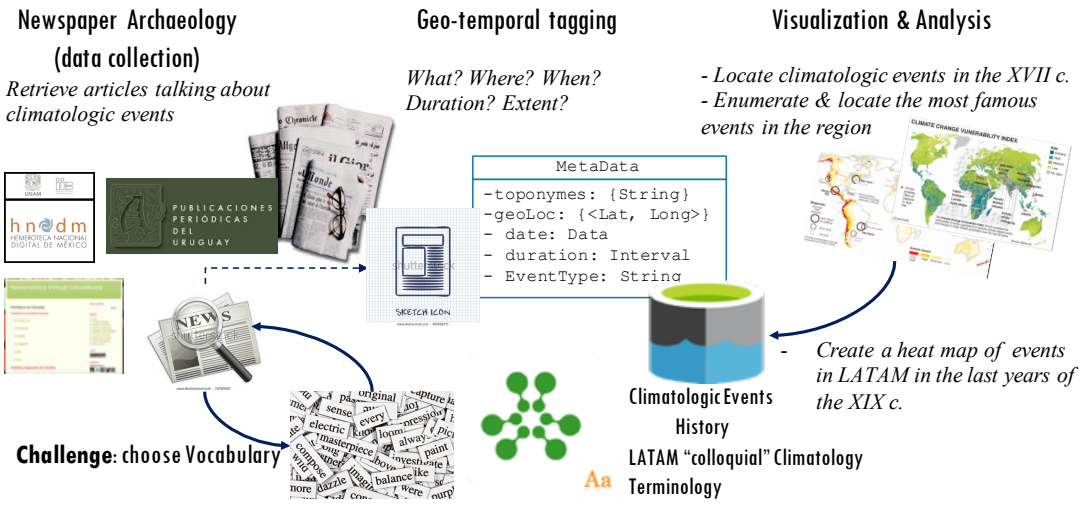

Fig. 1. Problem statement: building the history of climate change in Latin America between the XVIII and the XIX centuries

This information must be filtered, extracted of event deduced from the text.

The third aspect concerns visualization and analytics issues. The idea is to explore the climatologic events history built by identifying events during the process of exploring the articles found in the digital collections. The idea is to express analytics and relational queries and visualize them in cartographies on maps that can show different aspects of the event drawn in cartographic layers. For example, locate events happened during the XIX century, enumerate and locate the most famous climatological events in the region, create a heat map of the events in Latin America that happened in the last ten years of the XIX century.

To be sure that the answers can contain the maximum of items from the newspapers collections of the three four countries, it is important to explore the articles considering the language variations of the terms used for describing the events. The idea is to have a set of initial vocabularies using colloquial terms for describing climatologic events. Such vocabularies will constitute initial ontologies used for expressing and completing queries (synonym, related topics, etc.). These ontologies will be fed and enriched during the newspapers exploration process with the terms used in their content [2], [3].

Figure 2 shows that digital newspaper collections remain in the initial repositories that belong the libraries. Then, terms and links to the OCR (Optical Character Recognition) archives containing documents with articles reporting climatological events are stored in distributed histories managed in each country. Finally, local vocabularies are created, updated and enriched through queries, exploration and analytic activities. Users will decide to use some terms that can belong to any of these vocabularies, and then using query rewriting techniques these queries will be extended with synonyms, subsuming and general terms. The particular characteristic of these tasks is that the user (i.e., data analyst) can interact and guide the process according to her knowledge and expectations about what she expects to explore and search. This user guidance is called human in the loop.

This paper presents our approach for rewriting queries in order to better exploit data collections so that data scientists can be sure that they exploit as much data as possible.

3 Rewriting Queries for Exploring Climatologic Events

Queries as an answer is an data collections exploration technique that proposes rather than responding to a priori bad queries (too long to run, too many tuples), it gives a list of queries based on the information within a given data collection. The key challenge is identifying interesting queries to return. This could be done using a list of frequently used queries and returning them based on user feedback.

In the case of our project, data analysts express queries that can potentially explore newspapers content trying to have a good balance between precision and recall despite the ambiguity of the language (Spanish variations in naming climate events).

Our approach uses classic query rewriting strategies where given an initial conjunctive/disjunctive key word query using a colloquial vocabulary for denoting climate events the query is rewritten by extending it with general and more specific terms, synonyms, etc. The rewriting process can be automatic or interactive, in which case the system proposes alternatives and the user can validate the proposed terms. For example, if the query is *"heavy storms"*, the query can be completed by adding *"heavy storm trooper"*, *"heavy storm dust"*.

Furthermore, in our approach we also use knowledge domain information for rewriting the query, for example we have knowledge provided by experts stating that in the presence of a storm, the wind speed ranges between specific numbers, that the rivers can grow in some percentage depending on the litres of rain associated to the storm. Our approach uses this information for generating possible queries that can help the data scientist better precise her query, or define several queries that can be representative

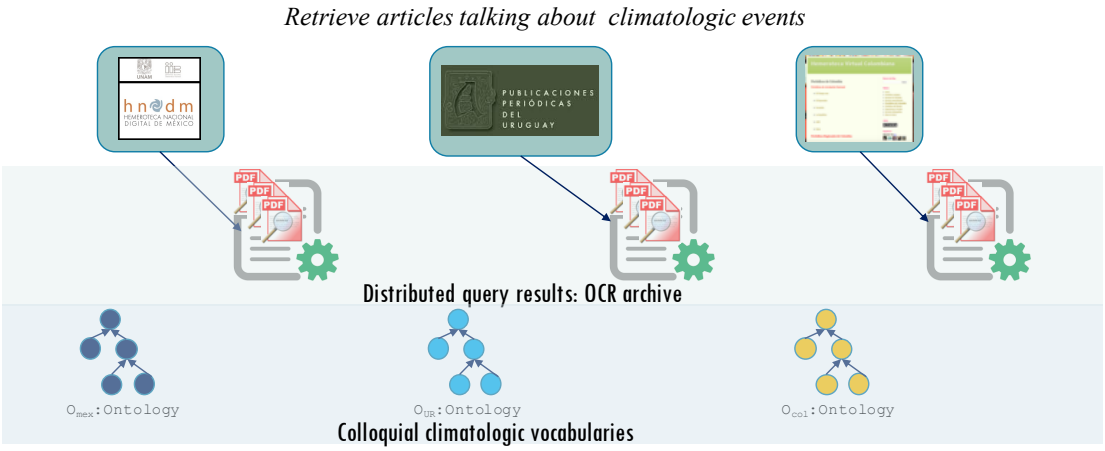

Fig. 2. Latin American climate change history exploration environment

of what she is looking for. For example, the previous initial query "Q_1: heavy storm" is rewritten into a new query "Q_{11}: heavy storm *or storm with 100 Km wind speed*". Since we can know using stored knowledge that the range of surface that can be reached by a 100 Km wind speed storm it is possible to estimate the surface of land reached by a storm. So other possible queries can look for "Q_{12}: storms with 100 Km speed that reached Mexico City". We can look for "Q_{13}: storms touching villages 500 Km around Mexico city happening in the same period". Instead of having a long query expression, our approach proposes sets of queries that the data scientist can choose and combine.

Query rewriting process is based on a "queries as an answer" process that uses:

- Wordnet for looking for associated terms and synonyms that help to address concepts used in different Spanish speaking countries. We do not translate the query terms to other languages because our digital data collections consist of newspapers written in Spanish.
- Climate glossary (Lode) that for a given term referring to a climatologic event, it has associated physical variables that characterise it. These information is used for generating new queries and also to complete a curated events base.
- A series of what we call "folksonomies" in a metaphorical way which are ontologies created through the processing of the vocabulary of newspaper articles. We create and feed each "folksonomy" according to the country of origin of the origin of the processed newspaper article. This let us extract the vocabulary used during the XVIII and XIX centuries for describing climatologic events in different Latin American countries (i.e. Mexico, Colombia, Ecuador and Uruguay).
- Digital newspapers collections with OCR's files and associated metadata describing the newspapers content like date, geographical location, name of the newspaper, pages, etc.
- Curated event store containing articles talking about already identified climatologic event.

Given a query expressed as a conjunction and disjunction of terms potentially belonging to a climate vocabulary, the query is represented in an expression tree where nodes are conjunction and disjunction operators and leafs are terms. The evaluation process of the query is performed first on top of the curated event store that leads to a set of a list of already curated events (see Figure 3).

Next we compute a list of queries that can be used for better targeting required information. We assume that a curation process has generated data structures that provide an abstract representation of the content of each article describing an event as shown in Figure 3. Classes of documents associated to an event (class Curated Event) with variables that describe its characteristics, of course the date in which it happened, the geographical scope. Using this information the following steps are performed for computing queries alternatives. For each leaf in the expression tree of the query:

- Use Wordnet [1] 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. The result is a new expression tree corresponding to an extended query Q_{ExT}.
- Use the climate glossary for transforming Q_{ExT} into queries with terms that can serve as filters. In the glossary there are variables concerning meteorology like wind speed, rain volume/hour, water level of sees, rivers and lakes. Other variables concern geographic aspects, like the location of an event, the scope of land it reaches. Finally, other variables concern damages caused by a climate event with specific physical and geographic characteristics. Our strategy is to generate queries combining variables of the same group, and of different groups. For example

1. http://timm.ujaen.es/recursos/spanish-wordnet-3-0/

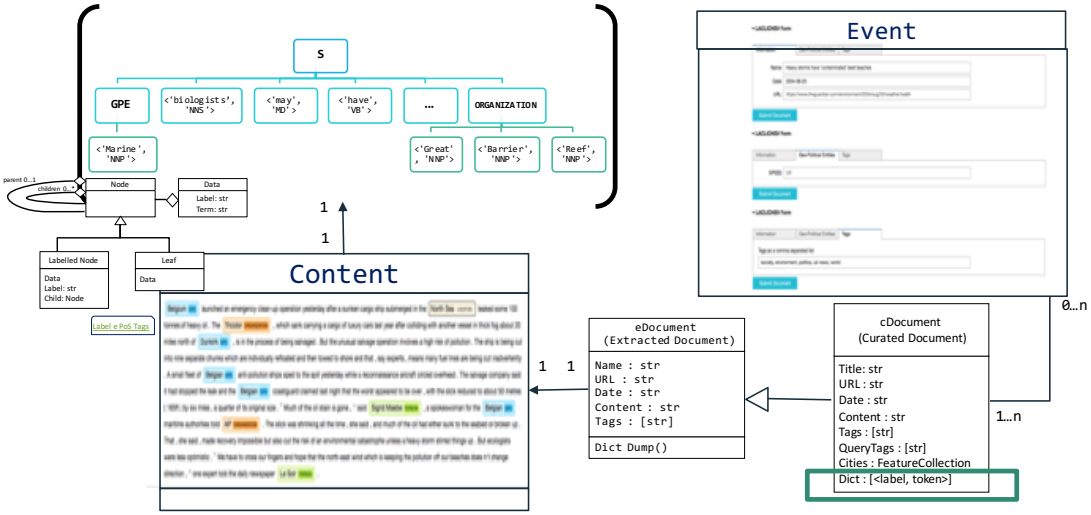

Fig. 3. Data structures describing the content of the articles talking about an event

"heavy storms with winds higher than 150 Km/h", "heavy storms with rains higher the 10 mm per square metre", "heavy storms with rivers' overflow". The result is a set of queries Q'$_{ExT1}$ ... Q'$_{ExTj}$.

- Use "folksonomies" for generating new query expression trees that substitute the terms used in Q'$_{ExTi}$ with equivalent terms used in a target country (e.g., blizzard instead of heavy storm). This will result in transformed expression trees each one using the terms of a country (Q"$_{ExT1}$ ... Q"$_{ExTj}$).

Implementation and experimentation The queries as an answer based rewriting process is illustrated in Figure 4. The "folksonomies" and the glossary have been specified using OWL and they are managed behind a service that is in charge of querying them and feeding them with new terms.

As seen in the diagram, there are managers devoted to query and maintain the glossary and folksonomies used by a query re-writer in charge of building the queries as an answer. We are currently designing and experiment in order to generate synthetic queries that can measure the cost of our initial strategy for generating queries that can potentially increase the possibilities of targeting data scientists expectations. The results with be assessed by measuring the degree of balance between precision and recall of every generated query.

This query rewriting service is part of the LACLICHEV environment developed in the context of the PAIGH project described in 2.

4 Related Work

Algorithms for rewriting queries have been proposed in two domains: (i) on the database domain; and (ii) on the service-oriented domain.

In the database domain, the query rewriting problem using views have been widely discussed [4], [5], [6], [7]. For instance, the *bucket algorithm* [5], *inverse-rules algorithm* [6] and *MiniCon algorithm* [7] have tackled the rewriting problem on the database domain.

Generally, data integration solutions on the service-oriented domain deal with query rewriting problems. [7] has inspired rewriting methods in service-oriented domain [8], [9]. [10] proposes a query rewriting approach which processes queries on data provider services. The query and data services are modelled as RDF views. A rewriting answer is a service composition in which the set of data service graphs fully satisfy the query graph. [11] introduces a service composition framework to answer preference queries. Two algorithms inspired on [10] are presented to rank the best rewritings based on previously computed scores. [9] extends [12] and presents an refinement algorithm based on *MiniCon* that produces and order rewritings according to user preferences and scores used to rank services that should be previously define by the user.

[13] proposes a query rewriting method for achieving RDF data integration using SPARQL. The principle of the approach is to rewrite the RDF graph pattern of the query using data manipulation functions in order to: (i) solve the entity co-reference problem which can lead to ineffective data integration; and (ii) exploit ontology alignments with a particular interest in data manipulation. The objective of the approach is: (i) solve the entity co-reference problem which can lead to ineffective data integration; and (ii) exploit ontology alignments with a particular interest in data manipulation.

[14] introduces a system (called SODIM) which combine data integration, service-oriented architecture and distributed processing. SODIM works on a pool of collaborative services and can process a large number of databases represented as web services.

In general, these approaches share the same performance problem as the traditional database algorithms. Furthermore, they do not take into consideration user's integration requirements which can lead to produce rewritings that are not satisfactory to the user in terms of quality requirements and cost.

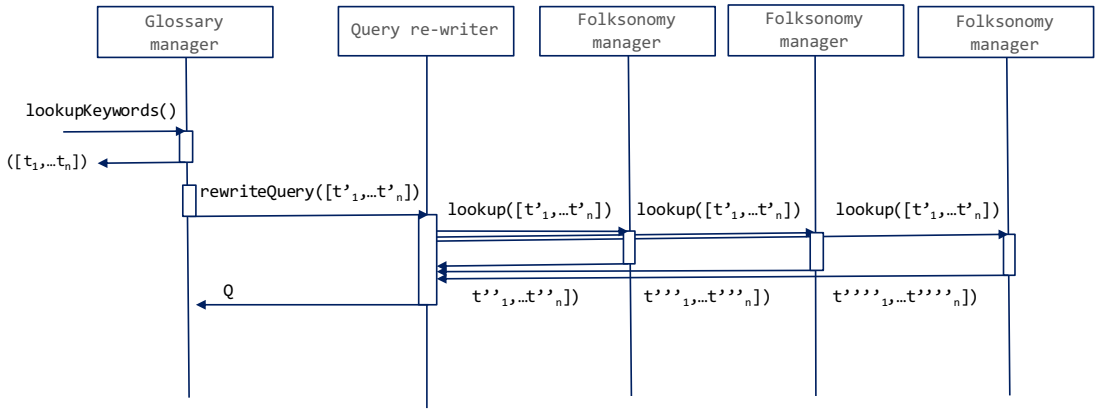

Fig. 4. UML sequence diagram of the general query rewriting process

5 Conclusion and Future Work

This paper introduced the first steps of our work regarding the generation of queries as an answer, given an initial query expressed for retrieving newspaper articles describing climate events that could have happened between the XVIII and XIX centuries in Mexico, Ecuador, Colombia and Uruguay. The approach uses knowledge about the vocabulary used in the content of the digital collections. This vocabulary is expressed as ontologies that we call "folksonomies" and that represent colloquial vocabulary used in each participating country for describing climate events. It also uses domain knowledge that can complete the description of events identified within newspapers articles but also for deriving more precise queries.

We are currently designing an experimental setting of our rewriting approach and integrating the service that implements it with the other services of an existing tool called LACLICHEV used for analysing newspapers articles for exploring climate change in Latin America during the XVIII and XIX centuries.

Acknowledgments

This work has been supported by the Panamerican Institute of Geography and History (PAIGH). Besides, we thank master student Estela Zamora Martínez of the Universidad de Guadalajara who is currently implementing a query recommendation service based on rewriting techniques funded by the CONACYT fellowship program of the Mexican government.

References

[1] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi, "Big data and its technical challenges," *Commun. ACM*, vol. 57, no. 7, pp. 86–94, Jul. 2014. [Online]. Available: http://doi.acm.org/10.1145/2611567

[2] U. Javeriana, "Ontoclima: Ontologa para representar eventos meteorolgicos registrados en la prensa." http://pegasus.javeriana.edu.co/ PA173-2-OntoClima/Entregables/OntoClima.owl.

[3] ——, "Ontologa individual meteocolombia." http://pegasus.javeriana.edu.co/ PA173-2-OntoClima/Entregables/MeteoColombia.owl.

[4] A. Y. Halevy, "Answering queries using views: A survey," *The VLDB Journal*, vol. 10, no. 4, pp. 270–294, Dec. 2001.

[5] A. Y. Levy, A. Rajaraman, and J. J. Ordille, "Querying heterogeneous information sources using source descriptions," in *Proceedings of the 22th International Conference on Very Large Data Bases*, San Francisco, CA, USA, 1996.

[6] O. M. Duschka and M. R. Genesereth, "Answering recursive queries using views," in *Proceedings of the Sixteenth Symposium on Principles of Database Systems*. NY, USA: ACM, 1997, pp. 109–116.

[7] R. Pottinger and A. Halevy, "Minicon: A scalable algorithm for answering queries using views," *The VLDB Journal*, pp. 182–198, 2001.

[8] U. Da Costa, M. Halfeld Ferrari, M. Musicante, and S. Robert, "Automatic Refinement of Service Compositions," in *ICWE*, F. Daniel, P. Dolog, and Q. Li, Eds., vol. 7977. Aalborg, Denmark: Springer, Jul. 2013, pp. 400–407. [Online]. Available: https://hal.inria.fr/hal-00861101

[9] C. Ba, U. Costa, M. H. Ferrari, R. Ferre, M. A. Musicante, V. Peralta, and S. Robert, "Preference-driven refinement of service compositions," in *Int. Conf. on Cloud Computing and Services Science*, ser. Proceedings of CLOSER 2014, 2014.

[10] M. Barhamgi, D. Benslimane, and B. Medjahed, "A query rewriting approach for web service composition," *Services Computing, IEEE Transactions on Services Computing*, 2010.

[11] K. Benouaret, D. Benslimane, A. Hadjali, and M. Barhamgi, "FuDoCS: A Web Service Composition System Based on Fuzzy Dominance for Preference Query Answering," 2011, 37th International Conference on Very Large Data Bases (VLDB 2011).

[12] U. Costa, M. Ferrari, M. Musicante, and S. Robert, "Automatic refinement of service compositions," in *Web Engineering*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2013, vol. 7977.

[13] G. Correndo, M. Salvadores, I. Millard, H. Glaser, and N. Shadbolt, "SPARQL query rewriting for implementing data integration over linked data," in *Proceedings of the 1st International Workshop on Data Semantics - DataSem '10*. New York, New York, USA: ACM Press, 2010.

[14] G. ElSheikh, M. Y. ElNainay, S. ElShehaby, and M. S. Abougabal, "SODIM: Service Oriented Data Integration based on MapReduce," *Alexandria Engineering Journal*, 2013.