Wearable Queries: Adapting Common Retrieval Needs to Data and Users (Vision Paper)

Barbara Catania, Giovanna Guerrini
University of Genoa, Italy
<name.surname>@unige.it

Alberto Belussi
University of Verona, Italy
alberto.belussi@univr.it

Federica Mandreoli, Riccardo Martoglia
University of Modena and Reggio Emilia, Italy
<name.surname>@unimore.it

Wilma Penzo
University of Bologna, Italy
wilma.penzo@unibo.it

ABSTRACT

The wealth of information generated by users interacting with the network and its applications is often under-utilized due to complications in accessing heterogeneous and dynamic data and retrieving relevant information from sources having possibly unknown formats and structures. Processing complex requests on such information sources can, thus, be costly, though not guaranteeing user satisfaction. Furthermore, dynamic contexts prevent substantial user involvement in the interpretation of the request.

The paper envisions an innovative solution to process the above mentioned requests, limiting user involvement by exploiting information on: (a) user context (geo-location, interests, needs); (b) data and processing quality; (c) similar requests repeated over time. By interpreting a request in a novel way by means of a Wearable Query (WQ), i.e., a query that captures the user and request specificities, we envision a methodological and technological solution for WQs in the presence of repeated information needs in distributed, heterogeneous, dynamic environments, with emphasis on the geo-spatial dimension and on data quality.

1. INTRODUCTION

User interactions with the network and its many applications generate a valuable amount of information, facts, and opinions with a great socio-economic potential. This huge wealth of information, however, is currently being exploited much below its potential because of the difficulties and limitations in accessing data to retrieve relevant information.

These difficulties concern the characteristics both of information sources and data they contain and of requests. On the one hand, data from different sources are highly heterogeneous in terms of structure, semantic richness, and quality. They are also increasingly geo-referenced, time-variant, and dynamic. Information sources may contain: strongly related and semantically complex but relatively static data (e.g., Linked Open Data); unstructured data, or data with a simple and defined structure; data dynamically generated by a multitude of diverse people (crowdsourced data); highly dynamic data generated by public or private institutions linked to the territory.

On the other hand, users would like to find answers to complex requests expressing relationships among the entities of their interest, but they are able to specify such requests only vaguely, because they cannot reasonably know format and structure of data encoding the relevant information. For example, the user may ask for the nearest shops selling the book which her friend Luca likes or the biography of the author of the painting she is watching.

Processing complex requests on heterogeneous and dynamic information sources can therefore be costly: the request must be interpreted, then processed on available sources deemed relevant, and finally the results aggregated in a consistent answer to be returned to the user, before the context changes. The generated answer may not, however, guarantee the user satisfaction, since it could be incorrectly interpreted, or could be processed on inaccurate, incomplete, unreliable data, or could finally require a processing time inadequate to the urgency of the request. The solution proposed in the literature, that is, a significant involvement of the user during the process of interpreting the request [7], does not seem adequate in dynamic contexts, in which the interaction is often hampered by the communication means and by the need to quickly obtain answers.

The paper envisages an approach for providing approximate answers to shared and complex information needs, even vaguely and imprecisely specified, operating on the full spectrum of relevant content, and thus overcoming the difficulties related to heterogeneity and dynamism, while requiring a limited user involvement.

The key element of the envisioned solution to overcome the limitations of current approaches is a new concept of adaptivity based on three coordinates relevant in this con-

text: (a) user profile and request context, with specific reference to geo-localization, (b) data and processing quality, (c) similar requests repeated over time. (a) and (b) are aspects known in the literature but they have never been considered in this context. However, as underlined in [10], with reference to the vastness, the origin, and the variety of content of interest, (a) allows us to overcome the logic of one-size-fits-all without overloading the user with useless results while (b) enables us to distinguish trustworthy sources from lower-quality ones and to explain the provenance of results. For example, thanks to (a) and (b), it is possible to choose the level of detail of an answer according to the user's background or to prefer synthetic and timely answers sacrificing the quality of result in the case of a user on the move or in an emergency situation. (c) is instead an unexplored aspect, though it is very common in dynamic contexts. It occurs, for example, during or after an exceptional event (environmental emergencies or flash mobbing initiatives), in the context of users belonging to the same community or that are in the same place, possibly at different times. The information needs are widespread among different users, because induced by the event, the interests of the community, and the place, respectively.

The ability to take advantage of the experience gained by prior processing in new searches allows response times and interpretation errors to be limited, thus reducing the possibility of producing unsatisfactory answers. In this way, the approach constitues a step towards the realization of an entity-relationship search paradigm for uncontrolled and wide information domains that will enable innovative and relevant applications, with an impact on qualitative and quantitative performance of systems for processing strongly interrelated and heterogeneous data in distributed dynamic environments.

An example of such application is a "Smart City Explorer" able to increase the users' knowledge of what characterizes the territory, taking into account the context, which varies dynamically, also in relation to the location of the users (e.g., current interests and needs, including the urgent need of a good/service, and the event/reason that generated the request) and their profile, more stable. Sources include Linked Open Data and social networks, as source of information (e.g., on flash mobbing initiatives) and comments. Information needs are shared because dependent on the location (e,g., in front of a monument) and/or a common need (e.g., bus schedule at a stop in case of cancellation of a train). The application selects in an adaptive way also the sources, the quality of which can vary depending on the geo-location of the data involved in the search.

Another type of application we envision is, for instance, to support post-disaster scenarios, characterized by the need to quickly "match" needs of goods and services with offerings (both geo-localized, specified as dynamic crowdsourced datasets and, in a vague and heterogeneous way, respectively) and to consult other open and semantically rich data to evaluate, e.g., roads, itineraries, and catalogs of alternative services. The information needs are shared by multiple users, as they depend on the event that caused the emergency.

Section 2 introduces the enabling concepts of Wearable Query (WQ) and Profiled Wearable Query Pattern (PWQP), providing an overview of the envisioned approach, while relationships with related work are discussed in Section 3.

2. OVERVIEW OF THE APPROACH

The solution is centered on the enabling concepts of Wearable Queries (WQ), which integrate explicit requests with profile and context, and Profiled Wearable Query Patterns (PWQP), that are synthetic representations of the correspondences between past WQs and sources used for their execution, associated with quantitative/qualitative effectiveness measures of the use of the PWQP at a certain time, with respect to certain contexts and/or profiles.

Queries are *wearable* with respect to user and request specificities, in terms of request context and user profile. Context information includes spatio-temporal coordinates of the request, its motivation, and its environment (e.g., in terms of potential interaction and urgency). The user profile includes information provided by the user (e.g., personal information) as well as induced by the system (e.g., user habits). It may include the user background and fields of interests so that data sources can be selected and results can be provided at the appropriate level of detail. A WQ is the request annotated with context and profile information.

Queries are however wearable also in that the provided results are devised keeping into account data specifities, and specifically the quality, the geo-localization, and the freshness of data sources and specific data items. With respect to quality and dynamism estimates, techniques proposed in the literature need to be enhanced to derive quality metadata from a comparison between reliable and inaccurate sources (e.g., crowdsourced ones), and to keep into account the geospatial dimension.

To enable search in a space of highly heterogeneous and poorly controlled sources, an adaptive pay-as-you-go approach influenced by quality, dynamics, and specificities of the considered sources needs to be adopted. The approach is incremental since it does not provide an integrated view, rather it generates and incrementally refines mappings between sources, according to the requirements induced by the submitted requests, thus avoiding the prohibitive costs of full integration. Approaches for mapping creation and refinement are needed that take advantage of information on the status of execution of the WQs generating them and of implicit and explicit feedback on the produced results. In the absence of a reference schema, associations between portions of entity-relationship queries and description of single sources, generated during processing, need to be maintaied.

The role of adaptivity in the approach is twofold: the space of sources is incrementally adapted to the peculiarities of the submitted requests and simultaneously requests are processed by incrementally adapting them to the peculiarities of the space of sources and its evolution over time.

For processing entity-relationship queries, we envisage a mechanism that moves at each step, in the large space of possible approximate answers, towards the sources deemed capable of producing the best solutions with respect to profile and context of the request, quality and dynamism of the sources and knowledge gained from previous executions. The process is incremental, i.e., first it attempts to exploit PWQPs, then makes a coarse-grain selection of sources, and later it focuses approximation efforts on the description of the selected sources. The results are composed or reconciled through mappings, selected or generated on-the-fly.

The envisioned solution couples this mechanism with a method for assessing the quality of each individual processing. This information, together with any explicit user feed-

back, is used for updates and refinements. Thus, PWQPs allow to capitalize on information gathered from previous processing of similar requests.

PWQPs are synthetic representations of a set of WQs processed in the past. A PWQP consists of a set of correspondences between query and data source portions, identified during past executions. The correspondences are associated with quantitative/qualitative measures about the effectiveness of using PWQPs at specific times, in relation to the contexts/profiles that generated them. These measures include estimates of both the mapping quality and the result quality (e.g., in terms of accuracy). For the former, the measures take into account the dynamic nature of the involved sources. The latter rely on the quality estimates associated with the data sources, the mappings used in the processing, and the adaptive process. Efficient techniques for estimating and refining these measures are needed. On the basis of such information, that is acquired, refined, and kept up-to-date during processing, and on its freshness with respect to the source dynamism, the PWQPs that best represent a WQ are selected, the rewriting of the WQ is performed, and the PWQPs are refined or new ones selected.

Solutions will be influenced both by dynamic and temporal aspects, and by the presence of textual content in crowdsourced, social, and micro-blogging data. The approximation techniques will adopt query expansion approaches by adding time constraints and targeted keyword-based searches, respectively. The mapping creation techniques will take into account the temporal dimension in the case of temporal mappings between temporal and/or dynamic sources.

The implementation of our vision involves achieving the following specific objectives and consequent results: (i) define models and languages for WQs and for the representation of source spaces; (ii) define a model to estimate the quality of the sources and processing, with particular reference to the quality of geo-referenced data; (iii) introduce techniques for managing PWQPs; (iv) introduce pay-as-you-go techniques for source spaces; (v) introduce adaptively approximate processing techniques for WQs.

3. RELATED WORK AND INNOVATION

The growing availability of knowledge bases, structured information from HTML sources, unstructured contents from social networks and microblogs, highlights the necessity of new paradigms for the representation and approximate querying of heterogeneous data that join structural and textual aspects, taking into account their spatiotemporal dimension that allows one to define the information context and identify their dynamicity [1]. Nowadays such data sources often include geo-spatial data [9] that are frequently provided by final users (e.g., through GPS tools). Data sources with this rate of heterogeneity are characterized by highly variable data quality, which is expressed by metadata that describe (at different granularity) data accuracy, completeness, consistency, and update rate and that can be profiled to get estimates on both data sources and query result [8]. Recent approaches also relies on explicit provenance information [12], describing data origin and the transformations they have undergone, for quality assessment. Metadata about result quality can be obtained through direct (or indirect) user feedback and used for iteratively refining the queries [7, 10]. Information on data dynamicity, and corresponding source variations, can be collected directly by the system [3].

Data fusion is performed by applying pay-as-you-go integration approaches [4] where the query processing is the result of rewriting techniques that apply mappings [2] and methods to detect entity matching [6, 11].

The concept of contextualized and repeated requests on dynamic sources, processed in terms of Wearable Queries, is innovative and originally reinterprets the notion of adaptive processing. Adaptive processing techniques [5] are primarily motivated by variable constraints on the availability of resources, also network ones (bandwidth, reliability).The proposed approach is novel in considering adaptivity in interpreting complex and shared information needs, and in processing the corresponding requests, by delegating to the system the choice of the best interpretation of repeated requests on the basis of information gathered in previous processing of similar requests. Adaptivity is therefore aimed primarily at improving the satisfaction of the user, also in terms of result quality. Another element of innovation is the mutual adaptivity of queries and sources.

The dynamic integration of crowdsourced data sources with more traditional and less dynamic sources is unexplored expecially for geo-localized sources, for which appropriate spatial data quality indicators still need to be devised.

4. REFERENCES

[1] S. Bedathur, K. Berberich, I. Patlakas, P. Triantafillou, and G. Weikum. D-Hive: Data Bees Pollinating RDF, Text, and Time. In *Proc. of CIDR 2013*, 2013.

[2] D. Calvanese, G. De Giacomo, M. Lenzerini, and M. Y. Vardi. Query Processing under GLAV Mappings for Relational and Graph Databases. In *Proc. of PVLDB 2013*, pages 61–72, 2013.

[3] J. Cho and H. Garcia-Molina. Estimating Frequency of Change. *ACM Trans. Internet Techn.*, 3(3):256–290, 2003.

[4] A. Das Sarma, X. Dong, and A. Halevy. Bootstrapping Pay-as-you-go Data Integration Systems. In *Proc. of SIGMOD 2008*, pages 861–874, 2008.

[5] A. Deshpande, Z. G. Ives, and V. Raman. Adaptive Query Processing. *Foundations and Trends in Databases*, 1(1):1–140, 2007.

[6] H. Köpcke and E. Rahm. Frameworks for Entity Matching: A Comparison. *Data Knowl. Eng.*, 69(2):197–210, 2010.

[7] Y. Mass, M. Ramanath, Y. Sagiv, and G. Weikum. IQ: The Case for Iterative Querying for Knowledge. In *Proc. of CIDR 2011*, pages 38–44, 2011.

[8] S. W. Sadiq, N. K. Yeganeh, and M. Indulska. 20 Years of Data Quality Research: Themes, Trends and Synergies. In *ADC*, pages 153–162, 2011.

[9] M. van Exel, E.Dias, and S.Fruijtier. The Impact of Crowdsourcing on Spatial Data Quality Indicators. In *Proc. of GIScience 2010*, 2010.

[10] G. Weikum. Data and Knowledge Discovery. Technical report, GRDI 2020 Scientific Paper, 2011.

[11] S. E. Whang, D. Marmaros, and H. Garcia-Molina. Pay-As-You-Go Entity Resolution. *IEEE Trans. Knowl. Data Eng.*, 25(5):1111–1124, 2013.

[12] J. Zhao and O. Hartig. Towards Interoperable Provenance Publication on the Linked Data Web. In *Proc. of LDOW 2012*, 2012.