(ICDE). Los Alamitos, California: IEEE Computer Society. ISBN: 978-1-5090-6543-1. http://dx.doi.org/10.1109/ICDE.2017.231

On Recommending Evolution Measures: A Human-aware Approach

Kostas Stefanidis
University of Tampere, Finland
kostas.stefanidis@uta.fi

Haridimos Kondylakis
ICS-FORTH, Greece
kondylak@ics.forth.gr

Georgia Troullinou
ICS-FORTH, Greece
troulin@ics.forth.gr

***Abstract*—As knowledge bases are constantly evolving, there is a clear need for monitoring and analyzing the changes that occur on them. Traditional approaches for studying the evolution of data focus on providing humans with deltas that include loads of information. In this work, we envision a processing model that recommends evolution measures taking into account particular challenges, such as relatedness, transparency, diversity, fairness and anonymity. We target at supporting humans with complementary measures that offer high-level overviews of the changes to help them understand how data of interest evolves.**

I. Introduction

Over the past decade, numerous knowledge bases, e.g., DBpedia, Freebase and YAGO, have been built to power large-scale knowledge sharing, but also an entity-centric Web search, mixing both structured data and text querying [1]. These knowledge bases offer comprehensive, machine-readable descriptions of a large variety of real-world entities (e.g., persons, places, products, events) published on the Web as Linked Data. Dynamicity is an indispensable part of the Linked Data [9]; such datasets are constantly evolving for several reasons, such as the inclusion of new experimental evidence or observations, or the correction of erroneous conceptualizations [17].

One natural way to understand how knowledge bases evolve, is to study their deltas between different versions. This has been proved to play a crucial role in various tasks, like the synchronization of autonomously developed dataset versions [2], the visualization of the evolution history of a dataset [10], the need for accessing previous versions of a dataset to support historical or cross-snapshot queries [13], and the integration [8] and synchronization [11] of interconnected Linked Data. Towards this direction, several approaches have been proposed for formally describing those deltas, ranging from low-level deltas (describing simple additions and deletions) to high-level deltas (describing complex updates, such as different change patterns in the subsumption hierarchy).

While both low-level and high-level deltas provide a descriptive analysis of the changes, none of them provide an overview of the changes or the parts of the knowledge bases that were mostly affected by the change process. To help humans get a supervisory overview of the changes, observe changes trends and identify the most changed parts of a knowledge base without requiring a significant amount of work from them, we target at recommending appropriate evolution measures that allow quantifying the changes that particular parts of a knowledge base underwent.

Specifically, we envision a general processing model, in which humans, who both generate and consume data, are in the core. That is, given that data is produced from several human activities, like on social networks, by sensors on the roads, or with online transactions, humans are really interested to be notified about how data evolve. Existing and additional evolution measures, flexible enough to capture the peculiarities and needs of different applications on dynamic data, can be exploited in order to suggest to humans different ways to understand and realize how their data evolve and which are the main changes, taking into consideration complementary viewpoints. For doing so, we propose to consider perspectives like relatedness, transparency, diversity, fairness and anonymity.

II. Exemplar Evolution Measures

In general, there are several ways for studying the evolution of a knowledge base[1]. Such ways can capture characteristics that are arguably important in order to quantify the intensity of the changes that a knowledge base underwent. So, for example, one can assume as important the amount (actual number) of changes that the classes or properties of a knowledge base underwent during the evolution process.

a) Number of class or property changes: Consider the evolution of a knowledge base from a version V_1 to a version V_2. In principle, low-level deltas can be used to describe the set of triples which were added ($\delta^+_{V_1,V_2}$) along with the set of triples which were deleted ($\delta^-_{V_1,V_2}$) during the evolution from V_1 to V_2. The number of detected changes over this evolution is the size of their low-level delta δ_{V_1,V_2}, i.e., $|\delta_{V_1,V_2}| = |\delta^+_{V_1,V_2}| + |\delta^-_{V_1,V_2}|$ [11]. For cases in which we are interested in the changes related to a particular class or property, we can specialize the above definition and take into account only the triples added and deleted referring to the class or property of interest. For instance, we can use $\delta_{V_1,V_2}(n)$ to denote the number of changes in which a class n appears.

b) Number of class or property changes in neighborhoods: Apart from the number of changes over a specific class or property, another interesting dimension is the number of changes in their neighborhoods. For example, when studying the evolution of a class n, we may be interested in the classes around n, thus allowing determining whether the topology of the knowledge base changed in a particular area. More specifically, for the scenario referring to the class n, we define its *neighborhood* for two versions V_1, V_2 (denoted by $N_{V_1,V_2}(n)$) as the set of classes that are either related to n via a subsumption relationship, or are connected with n via a

[1]For a preliminary study, see [16].

property (through the property's domain/range), in either of V_1, V_2. Then, the number of changes in $N_{V_1,V_2}(n)$ can be computed as: $|\delta^N_{V_1,V_2}(n)| = \sum_{c \in N_{V_1,V_2}(n)} |\delta_{V_1,V_2}(c)|$.

In a more sophisticated way, we can consider that the amount of interest related to a class or property is also related to how important this class or property is for the knowledge base, and how this importance changed during the evolution process. This importance can be captured by pure structural measures or by measures considering also semantics.

c) Structural Measures: The *Bridging Centrality* is a structural measure, which tries to identify the information flow and the topological locality of a node in a graph. A node with high Bridging Centrality is a node connecting densely connected components in a graph. Moving forward, the *Betweenness* of a class/node counts the number of the shortest paths from all nodes to all others that pass through that node. As such, a shift in one node's Bridging Centrality or Betweenness among V_1 and V_2 could capture how the different changes on an dataset affected the topology around this specific node.

d) Semantic Measures: The notion of *centrality* is used to quantify how central is a specific class in a specific version of a knowledge base [15]. To identify the centrality of a class n in a dataset version V_j, we calculate initially its *relative cardinality* by considering its corresponding instances. The relative cardinality $RC_{V_j}(e(n, n_i))$ of a property $e(n, n_i)$, which connects the classes n and n_i, is defined as the number of the specific instance connections between these two classes divided by the total number of the connections of the instances that the two classes have. Then, the data distribution is combined with the number of the incoming/outgoing properties of this class. As such, the in/out-centrality (C^{in}/C^{out}) is defined as the sum of the weighted relative cardinalities of the incoming/outgoing properties.

Relevance on the other hand, is a measure that extends centrality in order to consider neighborhoods as well. Intuitively, classes with many connections with other classes in the knowledge base should have a higher importance than classes with fewer connections. Thus, the relevance of a class is affected by the centrality of the class itself, as well as by the centrality of its neighboring classes. Moreover, since the knowledge base typically contains huge amounts of data, the actual data instances of the class are also considered when trying to estimate its importance.

Extensions on the above definitions can be given, so as to define the corresponding structural or semantic importance measures for properties as well. Having defined those measures, an indirect way of measuring the effects of a change on a class/propery is by determining how much the importance of that class/property has changed by means of the change in its importance measure - by computing the absolute difference of the importance measure before and after the changes. This is, in many cases, superior to the simple counting of changes, because it shows the cumulative effect of these changes on the class; and not all changes have the same effect. As such we can easily identify that there are many different views of evolution that we could consider according to the user's interest.

III. Humans in the Loop

When trying to identify the parts of a knowledge base that were mostly affected by the evolution process, we have to think about the humans that will use our results. Traditionally, curators, editors or groups of them are interested in understanding how data evolves and which are the most affected areas. Most often, to do that, we rely on measures, like the ones presented above, that allow quantifying the intensity of the changes that a piece of a knowledge base underwent.

Moving forward, our goal is to generalize that processing model and include players in the picture who generate data and, at the same time, are the targets of data analysis. Specifically, given that nowadays big data is produced from the human daily activities, e.g., on social networks, by sensors on the roads, or with online transactions, anyone at personal or group (e.g., family) level, may want to be notified about the evolution of data. Analyzing these diverse sources of data impacts the humans, who mostly were the makers of the data.

Overall, our focus in this line is to be able to recommend to the humans evolution measures or their mix that are qualified to cover different vertical and complementary viewpoints yet unexplored in such a setting. Next, we describe some of the issues in this loop.

a) Relatedness: Given the abundance of available information, exploring the contents of a knowledge base in order to study how it evolves is a complex process that may return a huge volume of data. Still, users would like to retrieve only a small piece of the evolved data, namely the most relevant to their interests and needs. In general, relevance is an important and well-studied criterion for ranking query results [6]. However, not enough work has been done towards associating the relatedness of the evolution of specific parts of a knowledge base with particular humans.

b) Transparency: Transparency helps humans to know what is being recorded for them and the evolution process, and how the recorded information is being used. Provenance information is important for achieving transparency, so that questions, such as who created this data item and when, by whom was the data item modified and when, and what was the processes used to create the data item, can be answered.

For such needs, usually workflow systems are employed. They support the automation of repetitive tasks, as well as they can capture complex analysis processes at various levels of detail and systematically capture provenance information for the derived data items. The provenance of a data item contains information about the process and data used to derive the data item. It provides important documentation that is key to preserving the data, to determining the data's quality and authorship, and to reproduce and validate the results [3], [12]. In our paradigm, provenance becomes important, since we care about the truth of the provenance data, taking into account the rules of a particular discipline. Data placed separately from its justification is meaningless, while for assessing its correctness and reliability three sources are typically used, namely, observation, inference and belief adoption.

c) Diversity: Recently, a big number of studies motivate the benefits diversity provides. Several definitions can be found in the research literature. Most of them can be

classified into [4]: (i) content-based, selecting data items that are dissimilar to each other, i.e., they do not contain overlapping information, (ii) novelty-based, selecting items that contain new information when compared to what was previously presented to the human, and (iii) semantic-based, selecting items that belong to different categories and topics. In all cases, diversity applies to a set of data items and not to individual ones.

In our case, the challenge is that we have to introduce algorithms resulting in sets of evolution measures that as a whole exhibit a desired property, and not assigning interest scores to measures individually. Namely, the produced set of measures should cover all the different needs of the human in question and not focus on a particular aspect of evolution. This problem becomes more difficult when we would like to locate the evolving parts of a knowledge base that a group of humans is interested in. This is a different aspect of diversity, because we cannot just combine the diverse measures produced for the humans in the group, since in this case we may construct a non diverse measures set.

d) Fairness: Abstractly speaking, fairness in data processing can be expressed as the lack of bias, where bias can come from data processing methods that reflect the preferences of the data scientists designing them [14]. Fairness, at the individual level, is hard to measure and to guarantee. In the sense of non-discrimination[2], the intuitive searching and ranking based on relevance is not enough, since, in that cases, we mostly care about common needs. Clearly, supporting uncommon information needs is important as well.

From a different perspective, the group notion of fairness can be handled by exploiting set-oriented ideas. For instance, assume that we would like to recommend evolution measures to a group of humans, e.g., the curators' team of a knowledge base. The goal here is to locate suggestions that include measures fair to the members of the group. Given a particular set of measures, it is possible to have a human u that is the least satisfied human in the group for all measures in the recommendations list, that is, all measures are not related to the interests of u. Therefore, although the group may like as a whole the set of recommendations, the package selection is not fair to u. In actual life, we should be able to recommend measures that are both strongly related and fair to the majority of the group members. Motivated by this observation, an important target is at having insights into the properties of the produced recommendations in order to help making the algorithmic process non-discriminative.

e) Anonymity: The typical process for observing the evolution of data is to find patterns that usually happen and perform some aggregations on them. Naturally, this is a method for achieving anonymity as well. For understanding why we need anonymity, consider a medical research scenario, in which the patient health records cannot be proceed individually because of their sensitiveness. Interestingly, data evolution can be studied from analyzing aggregations on them, thus sufficing privacy issues. But often, even if data is aggregated, it is possible to re-identify sensitive patient's data or significant parts of it [5]. Clearly, for many applications, like the ones in the health domain, access to personal and private data is essential, meaning that strict rules prohibiting reach such data should apply.

[2]Some observations about data-driven discrimination can be found in [7].

IV. Conclusions

In this work, we present a novel, recommenders-like way to assess the evolution intensity of knowledge bases. This is intended as an aid for the humans, allowing them to quickly understand how data changes and get an overview of the important changes under different perspectives.

Acknowledgments: This work was partially supported by the EU H2020 iManageCancer (#643529) project.

References

[1] V. Christophides, V. Efthymiou, and K. Stefanidis. *Entity Resolution in the Web of Data*. Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan & Claypool Publishers, 2015.

[2] R. Cloran and B. Irvin. Transmitting RDF graph deltas for a cheaper semantic Web. In *SATNAC*, 2005.

[3] S. B. Davidson, S. C. Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, and J. Freire. Provenance in scientific workflow systems. *IEEE Data Eng. Bull.*, 30(4):44–50, 2007.

[4] M. Drosou and E. Pitoura. Search result diversification. *SIGMOD Record*, 39(1):41–47, 2010.

[5] K. E. Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley. Research paper: A globally optimal k-anonymity method for the de-identification of health data. *JAMIA*, 16(5):670–682, 2009.

[6] I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Comput. Surv.*, 40(4), 2008.

[7] H. V. Jagadish. *Why We are Hard on Amazon and Should Be*, Aug 2016. Post at http://www.bigdatadialog.com/fairness/why-we-are-hard-on-amazon-and-should-be.

[8] H. Kondylakis and D. Plexousakis. Ontology evolution without tears. *J. Web Sem.*, 19:42–58, 2013.

[9] A. Mazeika, T. Tylenda, and G. Weikum. Entity timelines: visual analytics and named entity evolution. In *CIKM*, 2011.

[10] N. F. Noy, A. Chugh, W. Liu, and M. A. Musen. A framework for ontology evolution in collaborative environments. In *ISWC*, 2006.

[11] Y. Roussakis, I. Chrysakis, K. Stefanidis, G. Flouris, and Y. Stavrakas. A flexible framework for understanding the dynamics of evolving RDF datasets. In *ISWC*, 2015.

[12] Y. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Record*, 34(3):31–36, 2005.

[13] K. Stefanidis, I. Chrysakis, and G. Flouris. On designing archiving policies for evolving RDF datasets on the web. In *ER*, 2014.

[14] J. Stoyanovich, S. Abiteboul, and G. Miklau. Data responsibly: Fairness, neutrality and transparency in data analysis. In *EDBT*, pages 718–719, 2016.

[15] G. Troullinou, H. Kondylakis, E. Daskalaki, and D. Plexousakis. Ontology understanding without tears: The summarization approach. *Semantic Web*, 9(1):1–17, 2017.

[16] G. Troullinou, G. Roussakis, H. Kondylakis, K. Stefanidis, and G. Flouris. Understanding ontology evolution beyond deltas. In *EDBT/ICDT Workshops*, 2016.

[17] F. Zablith, G. Antoniou, M. d'Aquin, G. Flouris, H. Kondylakis, E. Motta, D. Plexousakis, and M. Sabou. Ontology evolution: a process-centric survey. *Knowl. Eng. Review*, 30(1):45–75, 2015.