Intelligent Nudging to Support Interactive Exploration of Big Data Graphs

Marwan Al-Tawil
School of Computing, University of Leeds, UK
scmata@leeds.ac.uk

ABSTRACT

This research investigates how to support the user exploration through big data graphs. Current successful approaches to interactive exploration take into account the utility from a user's point of view. In this PhD, we are focusing on *knowledge utility* – how useful the trajectories in a data graph are for expanding user's domain knowledge. The main goal of this research is to design intelligent nudging techniques to direct the user to 'good' trajectories for knowledge expansion. Our earlier work investigating empirical nudging strategies for users exploration, suggests that paths which start with familiar and highly inclusive entities and bring something unfamiliar are likely to increase the learning effect of users exploration. This direct us to investigate subsumption theory for meaningful learning and adopt it as our underpinning theoretical model to generate good trajectories, where familiar and highly inclusive entities are used as knowledge anchors to bring and learn new knowledge. This calls for developing an automatic approach to identify knowledge anchors in a data graph. We follow an analogy with basic level objects in domain taxonomies that underpin our automated approach for identifying knowledge anchors. Several metrics for extracting knowledge anchors in a data graph are developed and examined.

CCS Concepts

G.2.2 [Mathematics of Computing]: Graph Theory – Graph algorithms; H.1.2, 4.7 [Information systems]: Data management systems - Graph-based database models, WWW - RDF.

Keywords

Big Data graphs; interactive exploration; knowledge utility; knowledge anchors.

1. INTRODUCTION AND MOTIVATION

With the emergence and the growing rates of RDF Linked Data graphs, many applications take advantage of the exploration of the knowledge encoded in the graphs to support users' interactive exploration [3, 4, 17]. Consequently, more and more big data graphs are being exposed to users for exploratory search tasks such as learning or investigating, where the users usually discover new connections and associations [12]. Layman users who are engaged in exploratory search sessions will usually have no (or limited) familiarity with the specific domain and little (or no) awareness of the encoded knowledge in the graph. In other words, the *users' cognitive structures* about the domain may not match the *semantic structure of the data graph*. This can provide major obstacles to interactive exploration, especially when the users need to learn new things, resulting into confusion and frustration.

This research aims to support users' interactive exploration in big data graphs through directing the users to trajectories that can bring some benefit (utility) for the users (e.g. efficiency, effectiveness, motivation, knowledge expansion) [12, 19]. Specifically, we focus on *knowledge utility* – how useful a trajectory in a data graph is to expand one's knowledge in the domain. Earlier research has acknowledged that data graph exploration can promote expansion of domain knowledge through serendipitous learning (e.g. users discover concepts or relationships they were unaware of) [5, 15]. However, not all paths are beneficial for knowledge expansion, and ways for identifying 'good' trajectories are required.

Identifying good trajectories in a data graph requires anchoring entities that serve as knowledge bridges to learn new things. Our earlier work has acknowledged that when the user explores familiar entities, nudging should direct the user to explore something new [15]. This calls to investigate the subsumption theory for meaningful learning [6]. According to this theory, to incorporate new knowledge, the most familiar and inclusive entities in the user's cognition are used as *knowledge anchors* to subsume and learn new knowledge. Subsequently the new knowledge can take on meaning by becoming anchored with the basic concepts in the user's cognitive structures. However, identifying knowledge anchors in big data graphs is not a trivial task and brings forth various research challenges, including: dealing with larger number of entities, from 100s of entities in a typical ontology versus millions of entities in a typical data graph, and the need to exploit large number of data instances in the data graph.

The broader challenge of this PhD is: *to design intelligent nudging techniques to direct the user to 'good' trajectories through big data graphs for knowledge expansion*. To meet this challenge, we address two research questions:

Question 1: How to develop automatic ways to identify data graph entities that provide knowledge anchors for navigation paths? This question can be seen as focusing on the Cognitive Science notion of basic level objects[1] [7], to develop metrics to automatically identify knowledge anchors in a data graph.

Question 2: How can we use knowledge anchors in a data graph to design navigation paths for expanding users domain knowledge? In the second question, subsumption theories will be adopted to nudge the user through navigation paths. We aim to maximize the serendipitous learning through bringing the users first to the anchoring entities and then direct them to new and interesting concepts at different levels of abstraction in the graph.

2. RELATED WORK

Recent research on exploratory search through linked data graphs has been examining different ways to provide intelligent support for users' navigation. This has brought together research from the

[1] The term "basic level objects" has been used in Cognitive Science. Other developments, e.g. Formal Concept Analysts, call them "concepts.

Semantic Web, personalization, HCI and Cognitive Science to shape novel tools for interactive exploration of semantic data [14]. Work on personalized exploration includes improving search efficiency by considering user interests [8, 9, 17] or diversifying the user exploration paths with recommendations based on the browsing history [10]. Extracting semantic patterns from linked data sources to improve diversity in recommendation results to users has been proposed in [18]. Diversity is measured based on the semantic distance of topics and genres of the results. The work in [13] presents an approach to rank RDF statements with the expectation that some statements will be more valuable or interesting to users than other statements within some context. A wide range of tools for offering interactive exploration using linked data technologies can be found in a recent survey [12].

Our work brings a new dimension to this research stream by looking at the knowledge utility of the exploration path. We hypothesize that the cognitive learning theory of 'meaningful learning' [6] can be used to design paths with high knowledge utility, where new knowledge is subsumed under familiar and highly inclusive abstract entities. To identify knowledge anchors in a data graph, we operationalize the notion of basic level objects [7]. The problem of extracting important concepts from information spaces using the notion of basic level objects has been tackled by two approaches, ontology summarization [2, 11] and in Formal Concept Analysis (FCA) [22, 23]. These approaches utilize basic level objects with the aim of identifying key concepts to help *domain experts* in understanding and reengineering of an ontology or a concept lattice respectively.

In our work, we apply the notion of basic level objects in a data graph to identify *anchoring entities which are likely to be familiar to layman users who are not domain experts*. The formal framework that maps Rosch's definitions of basic level objects and cue validity [7] to data graphs is a major contribution of our work. We are unique in our use for these anchoring entities to support interactive exploration of a data graph.

3. PROPOSED APPROACH

We follow two approaches to address the two research questions in this work, respectively:

1. Develop automatic ways to identify data graph entities that can be used as knowledge anchors for navigation paths. We achieve this objective by adopting the notion of basic level objects which was introduced in Cognitive Science research [7], illustrating that domain taxonomies include category objects which are at the basic level of abstraction. Basic level categories "carry the most information, possess the highest category cue validity, and are, thus, the most differentiated from one another" [7]. We adopt two approaches to identify knowledge anchors: *distinctiveness* approach, that is based on the formal definition of cue validity, to identify the most differentiated basic categories whose attributes are associated amongst the category members but are not associated to members of other categories; and *homogeneity* approach to identify basic categories whose members share many attributes together. The homogeneity approach is complementary with the distinctiveness feature. A basic category object with high cue validity will have high number of entities common to its members.

2. Develop navigation paths using knowledge anchors. To achieve this objective we adopt subsumption strategies for meaningful learning [6]. Two subsumption strategies, the subordinate and the super-ordinate strategies, will be used [24]. On the one hand, subordinate strategy can be used to direct the user to explore unfamiliar members linked to anchoring entities in the data graph (i.e. nudge to explore subclass entities of an anchor). n the other hand, the super-ordinate strategy will be used when the anchoring entities are members of new unfamiliar and more inclusive entities (i.e. nudge to explore superclass entities of an anchor).

4. CURRENT OUTCOMES

We formally describe and implement the metrics and the corresponding algorithms for identifying *knowledge anchors* [1]. The metrics were implemented by running SPARQL queries over the MusicPinta data graph stored in a triplestore [5]. The performance of the algorithms is examined using benchmarking sets with basic level entities identified by humans, corresponding to the cognitive structures humans form on the part of the world represented in the data graph. A free-naming tasks based user study in the music domain using the MusicPinta data graph, was carried out to identify the benchmarking sets. This resulted in two such benchmarking sets, StrongAnchors set that includes entities closest to the human cognitive structures, and WeakAnchors set that includes entities people are likely to recognize when they are on the lower level abstraction in the graph. Based on quantitative and qualitative analysis, the strengths and limitations of each metrics are assessed, and a hybridization approach is proposed.

5. CONCLUSION AND FUTURE WORK

Interactive data exploration is becoming a key daily life activity. It involves exploring large amount of data and deciding where to go next in the graph. The success of big data graphs to support interactive exploration brings forth the challenge of building intelligent approaches to nudge the user through beneficial paths with high knowledge utility. This emphasizes the importance of identifying anchoring entities in the graph that can be used to subsume and learn new knowledge.

Moving forward, The immediate future work is to apply the metrics for identifying knowledge anchors in another domain. The INSPIRE[2] data graph about career options will be used to identify anchoring career points, that can be used in assisting the users in identifying paths that will be beneficial for expanding their awareness of their career options, including short or longer-term career paths. In the long run, we aim to develop nudging techniques using the subsumption strategies. An initial probing algorithm can be used to identify users familiarity [20] (i.e. use knowledge anchors to identify new or interesting entities for the users). Another option is to have a quick probing at every knowledge anchor during the exploration. A user study will be conducted to evaluate the navigation strategies. Random paths will be our base-line for evaluating the navigation strategies.

The impact of this work is not limited to support data graph exploration. It can be also applied to ontology summarization, where anchoring entities allow capturing a lay person's view of the domain. Also, knowledge anchors can be used to initiate a dialog to solve the 'cold start' problem in personalization and adaptation.

[2] INSPIRE is a system under development in Birkbeck, University of London, about career guidance domain, particularly career transitions.

6. REFERENCES

[1] Marwan Al-Tawil, Vania Dimitrova, Dhavalkumar Thakker and Brandon Bennett. Identifying Knowledge Anchors in a Data Graph. In HT2016, Halifax, Canada.

[2] Troullinou, G., Kondylakis, H., Daskalaki , E., Plexousakis , D. RDF Digest: Efficient Summarization of RDF/S KBs. In ESWC, 2015.

[3] Popov, I.O., Schraefel, M., Hall, W., Shadbolt, N. Connecting the Dots: A Multi-pivot approach to data exploration. In ISWC, 2011.

[4] Thellmann, K., Galkin, M., Orlandi, F., and Auer, S. LinkDaViz – Automatic Binding of Linked Data to Visualizations. In ISWC, 2015.

[5] Thakker, D., Dimitrova, V., Lau, L., Yang-Turner, F. & Despotakis, D. Assisting User Browsing over Linked Data: Requirements Elicitation with a User Study. *In ICWE 2013*

[6] Ausubel, D. A Subsumption Theory of Meaningful Verbal Learning and Retention. In Journal of General Psychology. Volume 66, Issue 2, 1962, pp. 213-224.

[7] Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., &Boyes-Braem, P. Basic objects in natural categories. Cognitive Psychology, 1976, 8, 382-439.

[8] Sah, M. & Wade, V. Personalized Concept-based Search and Exploration on the Web of Data using Results Categorization. In ESWC 2013.

[9] Rossel,O. Implemention of a “search and browse” scenario for the LinkedData. In IESD, 2014.

[10] Vocht1, et, al. A Visual Exploration Workflow as Enablerfor the Exploitation of Linked Open Data. In IESD, 2014.

[11] Peroni, S., Motta, E., d'Aquin, M. Identifying key concepts in an ontology through the integration of cognitive principles with statistical and topological measures. In ASWC, 2008.

[12] Marie, N., Gandon, F. Survey of linked data based exploration. In IESD@ISWC2014.

[13] Dean, M., Basu, P., Carterette, B., Partridge, C. and Hendler, J. What to Send First? A Study of Utility in the Semantic Web. In (LHD+SemQuant), 2012, @ ISWC2012..

[14] Waitelonis, J., Knuth, M., Wolf, L., Hercher, J., and H. Sack. The Path is the Destination-Enabling a New Search Paradigm with Linked Data. In LD in the Future Internet @ Future Internet Assembly, 2010.

[15] Al-Tawil, M., Thakker, D. and Dimitrova, V. Nudging to Expand User’s Domain Knowledge while Exploring Linked Data. In (IESD), 2014, @ ISWC2014.

[16] Tanaka, J., & Taylor, M. Object Categories and Expertise: Is the Basic Level in the Eye of the Beholder? Cognitive Psychology, 1991, 23, pp 457-482.

[17] Marie, N., Corby, O., Gandon, F. and Ribiere, M. Composite interests’ exploration thanks to on-the-fly linked data spreading activation. In Hypertext 2013.

[18] Maccatrozzo , V., Aroyo , L., Robert , R. Crowdsourced Evaluation of Semantic Patterns for Recommen-dations. In UMAP 2013. LBR.

[19] Nunes, T., Schwabe, D. Exploration of Semi-Structured Data Sources. In (IESD), 2014, @ ISWC2014.

[20] Al-Tawil, M., Dimitrova, V., Thakker, D. Using Basic Level Concepts in a Linked Data Graph to Detect User's Domain Familiarity. In UMAP2015, Dublin, Ireland.

[21] Thakker, D., Despotakis, D., Dimitrova, V., Lau, L., Brna, P. (2012). Taming digital traces for informal learning: A semantic-driven approach. In Proceedings of EC-TEL 2012.

[22] Belohlavek, R., Trnecka, M. Basic Level in Formal Concept Analysis: Interesting Concepts and Psychological Ramifications. In IJCAI 2013.

[23] Belohlavek, R., Trnecka, M. Basic level of concepts in formal concept analysis. In ICFCA, 2012, pp 28-44.

[24] Ausubel D., Novak, J., Hasian, H. Educational Psychology: a cognitive view - Rinehart Winston, New York, 1978.