Fabio Ciravegna
Maria-Esther Vidal (Eds.)

ISWC-DC 2015

The ISWC 2015 Doctoral Consortium

Bethlehem, Pennsylvania, October 12th, 2015
Proceedings

Editors' addresses:
Department of Computer Science
The University of Sheffield
Regent Court, 211 Portobello,
S1 4DP, Sheffield, UK
f.ciravegna@sheffield.ac.uk;

Department of Computer Science
Universidad Simón Bolívar
Valle de Sartenejas
Caracas 1086, Venezuela
mvidal@ldc.usb.ve

Preface

The ISWC Doctoral Consortium (ISWC-DC) is held in conjunction with the International Semantic Web Conference and provides Ph.D. students working on the Semantic Web area the opportunity to share and discuss their research problems in a critical but supportive environment. ISWC-DC is open to submissions of students at different stages of a Ph.D. program. Evaluation of the submissions is conducted according to the student level into the Ph.D. program.

This volume contains doctoral proposals accepted at ISWC-DC 2015, which was held in Bethlenhem, Pennsylvania on October 12th 2015. The call for papers of the ISWC-DC 2015 edition was open for three types of proposals: Early Stage, Middle Stage, and Final Stage Ph.D. students. All the proposals included in this volume went through a peer-review process that resulted out in the acceptance of 12 out of 24 submissions. All the accepted submissions were evaluated by three international senior reseachers in terms of: Motivation and Relevance of the Proposed Research; Analysis of the State-of-the-Art; Novelty of the Proposed Approach; Relevance of the Initial Results; Feasibility of the Empirical Evaluation Methodology; Analysis of Lessons Learned, Open Issues, and Future Directions; and Clarity and Presentation. Additionally, all the papers and their corresponding evaluations were meta-reviewed by two additional senior researchers. Comments of the reviewers and meta-reviewers were taken into account by the authors and included in the camera ready versions of the accepted papers.

Accepted doctoral proposals cover Linked Data management techniques, Linked Data mining, and Knowledge Discovery. For what we are proud to call a very exciting scientific event, we would like to warmly thank, in no particular order: the authors of the papers, the Program Committee, the ISWC Local Chairs and General Chair, and the supporting institutions. Furthermore, we thank EasyChair for providing the tools to create this proceedings. Without the effort of all the above, the ISWC-DC 2015 could not have been successful.

October, 2015
Bethlehem

Fabio Ciravegna
Maria-Esther Vidal

Doctoral Consortium Chairs

Fabio Ciravegna, University of Sheffield
Maria-Esther Vidal, Universidad Simón Bolívar, Venezuela

Program Committee

Harith Alani, The Open University
Oscar Corcho, Universidad Politecnica de Madrid
Philippe Cudré-Mauroux, University of Fribourg
Claudia Damato, Università degli Studi di Bari
Fabien Gandon, Inria
Pascal Hitzler, Wright State University
Lalana Kagal, Massachusetts Institute of Technology
Diana Maynard, University of Sheffield
Enrico Motta, The Open University
Natasha Noy, Google
Axel Polleres, Vienna University of Economics and Business - WU Wien
Guus Schreiber, VU University Amsterdam
Elena Simperl, University of Southampton
Ziqi Zhang, University of Sheffield

Contents

Iterative Query Refinement for Exploratory Search in Distributed Heterogeneous Linked Data

Laurens De Vocht

Multimedia Lab, Ghent University - iMinds,
Gaston Crommenlaan 8 bus 201, 9050 Ghent, Belgium
laurens.devocht@ugent.be

Abstract. Task-oriented search scenarios go beyond retrieving information when a one-time perception of search tasks is neither possible nor sufficient. Such scenarios typically need further investigation, navigation or understanding of the search results. Formulating a search query is particularly difficult in case of distributed Linked Data sources, because they have many different relationships and vocabularies. Since users cannot realistically construct their intended query correctly at the first attempt, they need an environment in which they can iteratively refine what they are searching for. Therefore, this PhD thesis proposes an adaptive set of techniques and implements them for use cases in academics, industry and government to measure the effect on the user experience. We show that the set of techniques facilitates web applications in fulfilling task-oriented searches more effectively and that user interaction with search results indeed gradually refines search queries.

1 Introduction

Typically, when users formulate search queries to find relevant content on the Web, they intend to address a single target source that needs to match their entire query. In cases when users want to discover and explore resources across multiple sources they need to repeat many sequences of search, check and rephrase until they have precisely refined their searches. The application of the Web of Data to search, makes it possible to extend basic keyword searches by describing the semantics of data and enables humans and machines to work together using controlled vocabularies. This enables distributing search tasks across datasets directly benefiting from a semantic description. Due to the high degree of mismatches between the structure of Linked Data and the variety in vocabularies across different sources, exploring distributed heterogeneous data sources is considered challenging.

Exploratory search covers a broader class of tasks than typical information retrieval where new information is sought in a bounded conceptual area rather than having a specific goal in mind. The users' demand to discover data across a variety of sources at once, requires searching facilities adaptive to their adjustments while they discover the data that were just put at their disposal. In general exploratory search describes either the problem context that motivates the search or the process by which the search is conducted [12]. This means that the users start from a vague but still goal-oriented defined information need and are able to refine their need upon the availability of new information to address it, with a mix of keyword look-up, expanding or rearranging the search context, filtering and analysis. Such queries will start simple but become more complicated as users get more and more familiar with the data after a while.

The general focus is the iterative exploration of linked data spread across different structural heterogeneous data sources. As there is no immediate suitable benchmark methodology for this model, it is necessary to rely on user-centered approaches and to develop reproducible automated machine approaches (using a gold standard). These approaches can be used to evaluate the application of the model in several prototypes which in turn allows us to observe how it enhances test-users search productivity and understanding of the data.

2 Motivating Examples

The generic methods and techniques developed in this PhD thesis find their application scenarios in various socio-economic relevant areas in academia, public sector and private sector. We give explain and motivate an example for each of the sectors.

2.1 Academia

Here the focus is bridging the walled garden of institutional repositories for 'Science 2.0'. Much research data and publications are publicly available online, not only via institutional repositories. The evolution of the Web to the Web 2.0 enabled a wide range of lay users via wikis, blogs and other content publishing platforms to become the main content providers. Combining information resources over the walls leads to a high degree of mismatches between vocabulary and data structure of the different sources [9]. Science 2.0 benefits from this exchange of information, however it is still challenge to explore these resources [16].

2.2 Public Sector

This example integrates application data from many local governments in reusable single purpose applications for 'Smart Cities'. If local governments keep developing (ad-hoc) data models and structures for this data over and over, it requires constant revising the model of available data while in fact not being able to cope with newer technologies and applications without heavily investing in new support infrastructure. For example, instead of making a street event organization application only for a single municipality, which outlines municipal services needed and permits required depending on the type of event, governments develop an event organization application usable for all municipalities in the region [6]. However, this is not trivial because it requires a lot of investments, approaches and ideas before finally coming to such an agreement.

2.3 Private Sector

In the last example, the goal is to embed data visualizations in industry search applications. In the industry, cases like those in the pharmacy-industry involve many partners in the development of a product (e.g new medicine). Every partner focuses on providing data for a different aspect such as the clinical trials, compounds and processes. It is thus complex to build systems that integrate and align this variety of data. Typically this data is very well structured or has high quality meta-data. Besides the pharmacy-industry, also the media and entertainment industry can benefit from such a framework. When recombining data from multimedia archives or social media for storytelling, new hidden relations and trends among existing sources could be discovered by properly describing and aligning them, enabling applications developers to design a whole range of interesting and entertaining applications and visualizations [19].

3 Challenges

Mostly direct querying approaches were tried and applications were often built around a limited set of supported SPARQL patterns. Furthermore, SPARQL queries are still hard for end users or even developers, despite GUIs and advanced query builders. Only in the last years vocabularies are getting streamlined and linked data is maturing. This leads to much more possibilities compared to traditional keyword search. Exploratory search in the front-end makes sense and transitioning from traditional web search and retrieval is changing. More and more web users and scenarios where exploratory search is beneficial appear (even though the paradigm is not new as such). The additional effort required for mapping, interlinking and maintaining data sources (i.e. as Linked Data), improves their re-usability and makes the methods and techniques for exploratory search immediately applicable. In the latter there are two scenarios: one where two data sources need be explored without interlinking them and the other where the effort is made: initial extra effort vs. reduced effort for implementing exploratory search.

3.1 Research Questions

We investigate how users find the information they need and gain insights about the data being under exploration through applications that enable them to interact with distributed heterogeneous data sources. The following questions is required to be addressed for attaining a set of techniques for exploratory search:

- *Can task execution be effectively facilitated by revealing relations between resources, i.e. adequately addressing the user's intent?*
- *To which degree does the additional interaction positively influence the relevance and precision of the search results?*
- *How does a justification of the presented results influence the user's certainty in getting closer to achieving the task's goal?*
- *How does the refinement of a search query gradually improve by interacting with its search results?*

It is relevant to measure if and how well agreeing on semantics proves to be useful in tackling these issues. Our approach and evaluation illustrates how to apply semantic paradigms for search, exploration and querying.

3.2 Hypotheses

Our research questions induce the following hypotheses:

- Interacting with the search results refines and improves the result set because interaction with the result set makes the information contained in the initial search query more specific, leading to more and more targeted queries.
- When exploring the data, indications such as facets, visualizations (charts, graphs etc.) reduce the number of steps to achieve a task's goal.
- Ordering of search results does not affect the search, neither in terms of steps needed, nor its precision.

4 State of the Art

Most of the works in literature about exploratory search, semantic search and distribution of queries across data sources deal with one or more aspects and are either focused on the front-end or the back-end. Typically they are limited to either a homogeneous dataset or they are purely focused on resolving the heterogeneity. In exploratory semantic search all these aspects need to be integrated. To the best of our knowledge there is no system that does all this. Nevertheless, one of the main contributions in this work is the distinct support for search scenarios where the revealed relation is one that the user was not aware of beforehand; besides describing methods and techniques for web developers and search applications on how to integrate exploratory search. However, there have been a lot of projects that cover multiple of these aspects playing an important role to make the whole work together. Therefore, we divide the related work section into two parts: (i) the front-end, *search interfaces*; and (ii) the back-end, *semantic search engines*. The opportunities lie in adaptive techniques applicable to combinations of different linked data sources covering the entire work-flow from back-end to front-end without denormalizing the semantics along the way.

4.1 Search Interfaces

The set of tools focus on revealing relationships between resources and exploring them. They contribute to distinct example solutions and implementations of adaptive and intelligent web-based systems [1]. During exploratory searches, it is likely that the problem context becomes better understood, allowing users to make more informed decisions about interaction or information use [20]. Rather than immediately jumping to the result, the observed advantages of searching by taking small steps include that it allowed users to specify less of their information need and provided a context in which to understand their results [17]. The mSpace framework and architecture as a platform to deploy lightweight Semantic Web applications which foreground associative interaction is one of first such interfaces [15] where data is not presented as a graph but in parallel tabs. It has been discussed that graphs are not always useful, even for tasks where they are supposed to support even though they are often chosen as a representation form for data in RDF [10].

4.2 Semantic Search

Recent developments demonstrate that Linked Data has arrived on the level of local governments, public services and their target user group: citizens. Initiatives such as the European Commission's "Interoperability Standards Agency" (ISA) [1] enforce the use of Linked Data and its data model RDF. Such data models are key for a formal semantic representation of data resources. Semantic search is one of the main motivations behind bootstrapping the Web into the Web of intelligent agents. Work on Semantic Web search engines like Hermes [18] closely relate to the main research question of our work. Such engines rely preliminary on keywords as a starting position for the definition and specification of queries but some also support more advanced querying capabilities, including basic SPARQL graph patterns. In general, the semantic matching frameworks within these semantic search engines reside on the approach of matching graph patterns against RDF data. This kind of semantic matching mechanism is also widely implemented by a range of RDF stores. Another alternative is Poweraqua [11], a query answering system but like ours it neither assumes that the user has any prior information about

[1] `http://ec.europa.eu/isa/`

the underlying semantic resources. Relation similarities are determined and triples are linked by expressing the input query as ontology concepts after identifying and mapping the terminology using a dedicated service. A system survey on Linked Data exploration systems [13] learned that massive use of linked data based exploratory search functionalities and systems constitutes an improvement for the evolving web search experience and this tendency is enhanced by the observation that users are getting more and more familiar with structured data in search through the major search engines. An interesting example here leverages the linked data richness to explore topics of interest through several perspectives over DBpedia [14].

5 Proposed Approach

Based on our experience in use cases in different domains (academia, industry and government) we identify and investigate a set of techniques for aligning and exploring data and verify that they are applicable in each of the domains. We generalize these techniques and iteratively refine them in an experimental setting where the data and queries are chosen carefully to highlight certain aspects (as depicted in the evaluation plan) to make the techniques applicable beyond the initial use cases we investigate. The goal is to optimize exploration techniques to the greatest extent. This involves detecting patterns in the data and defining a strategy for querying them accordingly, thereby balancing between common - and more rare queries fitting each scenario.

5.1 Definition

The techniques focus on generating views and abstractions, i.e. implement a query translation mechanism, accessible for end-users through services, and user interfaces. The other part focuses on aligning the data sources. Each of the use cases focuses on different aspect: The academic use case focuses on presenting the data to the users and turning them available for querying. The industry use cases implement translation techniques for the search tasks to queries. The government use case focuses on the semantic descriptions of the data to be able to query the data.

5.2 Implementation

We developed a semantic model for searching resources in the Web of Data developed for data related to scientific research (e.g. conferences, publications, researchers) [4] [7]. We implemented the model with current state-of-the art Web technologies and demonstrated it to end-users. The model uses research objects to represent the semantically modelled data to the end-users.

Our approach leverages RDF, and the annotated semantic graph by relying on the fact that the vocabularies used in them can be linked. similar data of different source can thus be described in using the same terms, making it possible to explore these sources with the same queries. The user interaction with the RDF datasets occurs through a set of interfaces. Each interface facilitates the reuse, exposure and publication of digital research content as Linked Data. The interfaces bridge each of the components in the search infrastructure.

6 Evaluation Methodology

We elaborate on the evaluation methods and present intermediate results indicating the feasibility, effectiveness and usefulness of the techniques:

– **Case by case**: the evaluation focuses on the use cases overall user perception and information retrieval quality (Eeffectiveness). Thereby we are testing both the

(task-oriented) user experience and information retrieval aspects of each approach. We deduct as much as possible information out of these real-world proof of concept settings to address the research questions and hypotheses in.

- **Generic applicability**: each hypotheses is evaluated directly and each of the research questions is address individually, in a perfect environment. Individual aspects are to be tested on a standardized collection and a standardized set of queries, changing only a single parameter to be able to test the hypotheses. Specifically we want to test the effects of returning results as a set rather than a list; test where two data-sources are being explored without interlinking them and the other where the effort is made; and the impact on the number of steps or time needed to complete a task when justifications are presented and cases when they aren't.

In each of both cases, the approach is evaluated in two ways: (i) automated - by machines - after defining a suitable baseline for comparison (quantitative); (ii) user tasks - by observing user interactions with prototypes that implement the techniques (quantitative) and an accompanying user questionnaire (qualitative). Since the main purpose of the techniques is to facilitate users in exploring Linked Data on the Web, the evaluation of our approach is focused on both the end-users and the precision of the search results, as perceived by them.

Therefore, we investigate and define:

- the characteristics, worth to be evaluated, of the data used in the experiments and
- the baseline against which the search engine is evaluated.

Hereby the focus lies on information retrieval (IR) aspects which are important to quantify because it is inherent to any type of search (thus also exploratory search) and user-centered aspects. IR measures do not give the whole picture in exploratory search as they do in traditional query-centric search, in particular task-oriented, user centric, measures, are particularly useful evaluation criteria in exploratory search.

7 Intermediate Results

The processing of queries and mapping of keyword queries proved to be of promising precision, given the complex and dynamic nature of the used datasets: a combination of Linked Open Data and non Linked-data sources. We observed that searching by keywords for resources increases the result set with more new relevant resources, while it is on average as precise as expanding existing resources in the result set. The results of a short survey[8] indicated that end-users embrace and understand the main goals of approach using the prototype we have developed.

The final interface, provided to the end-users, gave abundant and accurate information about users, when the quality of the underlying alignment between datasets has high accuracy and minimum sensitivity [5]. Furthermore we evaluated aligned and interlinked user profiles with Linked Open Data from DBLP[2] and COLINDA[3] [16] and measured a relatively high accuracy when detecting conferences in tags and a promising sensitivity when interlinking articles and authors [5]. This achievement is essential for the effective realization of a tool to facilitate the personalized exploration of heterogeneous data sources containing both research data and social data. Both providers of research data will benefit, by opening up their data to a broader audience, and users, through actively using collaboration tools and social media.

Considering that the implementation is still in the prototype phase, the potential of a set of techniques to support visual and interactive search is well demonstrated and

[2] `http://www.informatik.uni-trier.de/~ley/db/`
[3] `http://wwww.colinda.org`

understood by the target users. This relies mainly on the generic algorithm we developed for revealing relations between Linked Data resources [2] It proves that the dynamic alignment of resources is useful for our set of techniques when it operates as the back-end for a visualization tool like ResXplorer[4], a radial graph interface for researchers [3]. Such applications make optimal use of our set of techniques and visualize the aligned profiles and resources to allow the exploration of the underlying research data.

8 Conclusions

We aim to deliver the core building blocks for user oriented search engines and to facilitate exploring Linked Data, and ensuring their effectiveness by measuring: (i) the search precision; (ii) the support for re-usability of underlying data; and (iii) the degree of which they make search task execution more efficient. This PhD thesis investigates methods and techniques for web applications to support iterative refinement of queries for exploratory search with Linked Data. Overall, supporting such exploration on top of Linked Data: turns the potential of its exploitation more likely; and while allowing a larger group of users to discover Linked Data at the same time it increases the demand for this type of data, both in terms of context and semantics.

The enrichment of the main used data sources with Linked (Open) Data sources allows users to find a vast amount of resources implicitly related to them and thus initially not accessible. Facilitating exploration and search across semantically described distributed heterogeneous data sources is useful because it is still a laborious task for users to construct separate search queries for each of those services separately. We show how end-user applications facilitate accurately and iteratively exploring of linked data, without the need for a traditional ranked list of results. The set of techniques contributes to authenticity of the data it models and processes by guaranteeing that the final output towards the user has useful results in its domain of application. Because we stick with our approach close to the original structure of the data, this method is applicable to other domains if it is adequately structured by adapting the chosen vocabularies according to the datasets used.

The techniques contribute to users desiring to iteratively formulate precise searches and discovering new leads or validating existing finding across heterogeneous data without having to hassle with trial and error using traditional search engines. This will allow links to be revealed available but also to incorporate network structured data such as social and research data beyond the typical single user's scope. This should lead to more fine-grained details facilitating users to obtain a more sophisticated selection and linking of contributed resources based on previous assessments and explored links.

Acknowledgments.

I would like to thank my supervisors: Ruben Verborgh, Erik Mannens and Rik Van de Walle, for their support and the opportunity for the realization of this work.

References

1. Brusilovsky, P.: Methods and techniques of adaptive hypermedia. In: Adaptive hypertext and hypermedia, pp. 1–43. Springer (1998)
2. De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Discovering meaningful connections between resources in the web of data. In: Proceedings of the 6th Workshop on Linked Data on the Web (LDOW2013) (2013)
3. De Vocht, L., Mannens, E., Van de Walle, R., Softic, S., Ebner, M.: A search interface for researchers to explore affinities in a linked data knowledge base. In: Proceedings of the 12th International Semantic Web Conference Posters & Demonstrations Track. pp. 21–24. CEUR-WS (2013)
4. De Vocht, L., Softic, S., Ebner, M., Mühlburger, H.: Semantically driven social data aggregation interfaces for research 2.0. In: Proceedings of the 11th International Conference on

[4] `http://www.resxplorer.org`

Knowledge Management and Knowledge Technologies. pp. 43:1–43:9. i-KNOW '11, ACM, New York, NY, USA (2011)
5. De Vocht, L., Softic, S., Mannens, E., Ebner, M., Van de Walle, R.: Aligning web collaboration tools with research data for scholars. In: Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion. pp. 1203–1208. WWW Companion '14, Republic and Canton of Geneva, Switzerland (2014)
6. De Vocht, L., Van Compernolle, M., Dimou, A., Colpaert, P., Verborgh, R., Mannens, E., Mechant, P., Van de Walle, R.: Converging on semantics to ensure local government data reuse. In: Proceedings of the 5th workshop on Semantics for Smarter Cities (SSC14), 13th International Semantic Web Conference (ISWC) (2014)
7. De Vocht, L., Van Deursen, D., Mannens, E., Van de Walle, R.: A semantic approach to cross-disciplinary research collaboration. iJET 7(S2), 22–30 (2012)
8. Dimou, A., De Vocht, L., Van Compernolle, M., Mannens, E., Mechant, P., Van de Walle, R.: A visual workflow to explore the web of data for scholars (2014)
9. Herzig, D.M., Tran, T.: Heterogeneous web data search using relevance-based on the fly data integration. In: Mille, A., Gandon, F.L., Misselis, J., Rabinovich, M., Staab, S. (eds.) WWW. pp. 141–150. ACM (2012)
10. Karger, D., et al.: The pathetic fallacy of RDF (2006)
11. Lopez, V., Motta, E., Uren, V.: Poweraqua: Fishing the semantic web. Springer (2006)
12. Marchionini, G.: Exploratory search: from finding to understanding. Commun. ACM 49(4), 41–46 (Apr 2006)
13. Marie, N., Gandon, F.L.: Survey of linked data based exploration systems. In: Proceedings of the 3rd International Workshop on Intelligent Exploration of Semantic Data (IESD 2014) co-located with the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy, October 20, 2014. (2014)
14. Marie, N., Gandon, F.L., Giboin, A., Palagi, É.: Exploratory search on topics through different perspectives with dbpedia. In: Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014. pp. 45–52 (2014)
15. schraefel, m.c., Smith, D.A., Owens, A., Russell, A., Harris, C., Wilson, M.: The evolving mspace platform: Leveraging the semantic web on the trail of the memex. In: Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia. pp. 174–183. HYPERTEXT '05, ACM, New York, NY, USA (2005)
16. Softic, S., De Vocht, L., Mannens, E., Ebner, M., Van de Walle, R.: COLINDA: Modeling, Representing and Using Scientific Events in the Web of Data. In: Proceedings of the 4th International Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2015) Co-located with ESWC 2015. pp. 12–23 (2015)
17. Teevan, J., Alvarado, C., Ackerman, M.S., Karger, D.R.: The perfect search engine is not enough: a study of orienteering behavior in directed search. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 415–422. ACM (2004)
18. Tran, T., Wang, H., Haase, P.: Hermes: Dataweb search on a pay-as-you-go integration infrastructure. Web Semantics: Science, Services and Agents on the World Wide Web 7(3) (2009)
19. Vander Sande, M., Verborgh, R., Coppens, S., De Nies, T., Debevere, P., De Vocht, L., Potter, P.D., Deursen, D.V., Mannens, E., Van de Walle, R.: Everything is connected: Using Linked Data for multimedia narration of connections between concepts. In: International Semantic Web Conference (Posters & Demos). vol. 914 (2012)
20. White, R.W., Roth, R.A.: Exploratory search: Beyond the query-response paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services 1(1), 1–98 (2009)

Peer-based query rewriting in SPARQL for semantic integration of Linked Data

Mirko Michele Dimartino

Birkbeck, University of London
mirko@dcs.bbk.ac.uk

Abstract. In this proposal we address the problem of ontology-based SPARQL query answering over distributed Linked Data sources, where the ontology is given by conjunctive mappings between the source schemas in a peer-to-peer fashion and by equality constraints between constants. In our setting, the data is not materialised in a single datastore: it is accessed in a distributed environment through SPARQL endpoints. We aim to achieve query answering by generating the *perfect rewriting* of the original query and then processing the rewritten query over distributed SPARQL endpoints. We identify a subset of ontology constraints that enjoy the *first-order rewritability* property and we perform preliminary empirical evaluation taking into account such restricted constraints only. For future work, we aim to tackle the query answering problem in the general case.

1 Scene Setting

The Web of Linked Open Data (LOD) has developed from a few datasets in 2007 into a large data space containing billions of RDF triples published and stored in hundreds of independent datasets. This huge information cloud, ranging over a wide set of data domains, poses a great challenge when it comes to reconciling heterogeneous schemas or vocabularies adopted by data publishers. According to Linked Data best practices [10], data publishers should reuse terms from widely-used vocabularies already present in the cloud, in order to enable the discovery of additional data and to support the integration of data from multiple sources. However, commonly-used vocabularies usually do not provide all terms needed to completely describe the content of the data. Thus, data providers often define proprietary terms as sets of new IRIs published on the cloud. This trend leads to the formation of islands of data describing overlapping domains, rather than generating a global knowledge base.

Over the past years, researchers in the Semantic Web community have attempted to tackle these challenges by proposing several approaches based on semantics preserving SPARQL rewriting algorithms. These methods allow users to pose SPARQL queries expressed adopting a preferred vocabulary and a rewriting algorithm provides translations of the query in the language of similar vocabularies. To rewrite queries, they reason over semantic mappings between the sources. Following this, the rewritten query is evaluated over the sources and a more complete answer is returned to the user. Several approaches in the literature address this problem, however most of these works are based on the common two-tiered local-to-global schema integration paradigm, where

query are expressed over the global schema and are reformulated in the language of the source vocabulary, to be then evaluated over the data stored at the source. By contrast, in the Linked Data cloud each data store is an autonomous system whose vocabulary should represent part of the global schema, available from a distributed environment. In this regard, Linked Data consumers should be able to pose queries adopting any of the source vocabularies and to access similar sources through query translation, in a transparent way and without relying on a single global schema. Following this idea, we believe that a peer-to-peer approach is more suitable than a local-to-global approach because it provides a more decentralised architecture where peers act both as clients and as servers during the query reformulation process. In addition, the local-to-global approaches typically require a comprehensive global schema design before they can be used, thus they are difficult to scale because schema evolution may break backwards compatibility. Scalability is a key property of LOD-oriented data mediation systems, due to the continuous increase of data published on the web.

In this proposal paper, we address the problem of ontology-based SPARQL query answering via query rewriting over Linked Data sources, where the ontology is given by conjunctive mappings between the source schemas in a peer-to-peer fashion, and by equality constraints between constants for entity resolution. We focus on a setting where data is not materialised in a single datastore; it is accessed through distributed SPARQL endpoints. Query answering is achieved via: *(i)* computing the *perfect rewriting* with respect to the ontology; the original query is reformulated so as to obtain a sound and complete answer based both on the extensional database (i.e., the stored RDF triples) and the ontology defined as the set of data constraints entailed by the mapping assertions; *(ii)* processing the perfect rewriting of the original query over distributed SPARQL endpoints. Differently from other approaches which focus on tractable mapping languages, we aim to tackle query answering based on conjunctive mappings, i.e., positive rules in which the conclusion may contain existentially quantified variables, which makes reasoning tasks undecidable if interpreted in first order semantics. In addition, we take into account the distributed nature of the LOD cloud by processing the queries directly over the sources through their public SPARQL endpoints, without the need to materialise the RDF sources in a centralised middleware.

This PhD thesis aims to address the following research questions:
RQ1 How can we achieve ontology-based SPARQL query answering where the ontology comprises conjunctive mappings, i.e., existential rules which lead to undecidability of reasoning tasks?
RQ2 How can we compute the perfect rewriting of conjunctive SPARQL queries with respect to conjunctive mappings (interpreted so as to preserve decidabilty) and equality constraints between RDF sources? Which query language is suitable for such rewritings?
RQ3 How do we process the rewritten query taking into account that the RDF sources are not materialised is a single datastore, and data is accessed via distributed SPARQL endpoints?

Related work. Several works in the literature address data mediation for Linked Data. Very close to our work is [5] which proposes an algorithm to rewrite SPARQL queries in order to achieve integration of RDF databases. The approach is based on

the encoding of rewriting rules for RDF patterns that constitute part of the structure of a SPARQL query. The adopted rules, called *Entity Alignments*, express semantic mappings between two datasets and can be interpreted as definite Horn clauses in First-Order (FO) logic where only the *triple* predicate is used. It is then based on the well-known *global-as-view* [13] data integration approach, where a term of the global schema is mapped to a view of the source. The main limitation of the proposed approach is that it is not applicable when a more expressive formalism is needed to align two schemas, for example, when the relations in the sources need to be specified as views over the mediated schema. In this scenario, the expressive power of the *local-as-view* [13] formalization is also necessary. One interesting aspect is that the framework deals with co-reference resolution by including *Functional Dependencies* in the mapping rules. Similar limitations are in the approach proposed by Makris et al. [15, 16] which uses a mapping language based on Description Logics, defining 1:N cardinality mapping types where a term from one vocabulary is mapped to a Description Logic expression over another vocabulary. Other similar approaches leveraging less expressive formalisms for data mediation can be found in [17, 20, 21]. For instance, [17] proposes SemLAV, an alternative technique to process SPARQL queries without generating rewritings. SemLAV executes the query against a partial instance of the global schema which is built on-the-fly with data from the relevant views. Work in [20] addresses rewriting techniques that consider only co-reference resolution in the rewriting process, and [21] adopts a small set of mapping axioms defined only by those RDF triples whose predicate is one of the following OWL or RDFS terms: `sameAs`, `subClassOf`, `subPropertyOf`, `equivalentClass`, and `equivalentProperty`. Other similar approaches are proposed in [12, 14, 19]. All the above-mentioned frameworks address query answering over two-tiered architectures and tractable mapping languages, while we wish to explore more general settings.

Several peer-to-peer systems for RDF datasources can be found in the literature. For instance, in [2, 3] the authors describe a distributed RDF metadata storage, querying and subscription service, as a structured P2P network. Similarly, work in [18] proposes routing strategies for RDF-based P2P networks. These are non-database-oriented tools that have little support for semantic integration of highly heterogeneous data. In fact, they focus strictly on handling semantic-free requests which limits their utility in establishing complex links between peers.

2 Proposed Approach

Our approach to semantic integration of heterogeneous Linked Data sources is based on the *RDF Peer System* (RPS) introduced in our recent paper [7]. This is a framework for peer-based integration of RDF datasets, where the semantic relationships between data at different peers are expressed through mappings. Formally, an RPS $\mathcal{P}$ is defined as a tuple $\mathcal{P} = (\mathcal{S}, G, E)$, where $\mathcal{S}$ is the set of the *peer schemas* in $\mathcal{P}$, G is a set of *graph mapping assertions* and E is a set of *equivalence mappings*. A peer schema in $\mathcal{S}$ represents the adopted vocabulary, that is, the set of IRIs that a peer (i.e. an RDF data source) adopts to describe its data. The sets of schema-level mappings and instance-level mappings between peers are given by G and E, respectively. G provides semantic linkage

between the schemas of different peers and contains mapping assertions of the form $Q \rightsquigarrow Q'$, where Q and Q' are conjunctive SPARQL queries with the same arity over two peers, e.g.: $q(x, y) \leftarrow (x, actor, y) \rightsquigarrow q(x, y) \leftarrow (x, starring, z)$ AND $(z, artist, y)$, where the query $q(x, y) \leftarrow (x, pred, y)$ evaluated over an RDF database returns the subjects and objects appearing on all the triples whose predicate is $pred$. For instance, this mapping assertion states that, if there is a triple in the first source of the form $(IRI_1, actor, IRI_2)$, then the two triples $(IRI_1, starring, \_{:}b)$ and $(\_{:}b, artist, IRI_2)$ need to be exported to the other source, where $\_{:}b$ is a blank node. Mappings in E are of the form $c \equiv_e c'$, where c and c' are IRIs located in the same peer or in two different peers. Equivalence mappings are used to solve the problem of identity in the Semantic Web scenario. In fact, an IRI ensures to uniquely identify a resource on the web, not the entity the resource represents [9]. To partially cope with this, LOD publishers often use the built-in OWL property `sameAs`[1], to explicitly "align" the newly created IRIs with existing IRIs that represent the same real-world entity. Equivalence mappings entail the semantics of `sameAs`.

Query processing in our setting is performed by query rewriting; the original query is reformulated so as to obtain a sound and complete answer based both on the extensional database (i.e., the stored RDF triples) and the ontology defined as the set of data constraints entailed by the mapping assertions. Different from the existing SPARQL rewriting approaches, our integration framework preserves the expressive power of both the global-as-view and local-as-view integration formalisms. In addition, we adopt a more general network of interrelated peer-to-peer relations. To the best of our knowledge, none of the existing approaches addresses SPARQL query answering under such a setting.

For the theoretical evaluations, we formalise the query answering problem by generalising the notion of *certain answers* [1] to our context. We show that the problem is subsumed by conjunctive query answering in data exchange for the relational model, and that a conjunctive SPARQL query can be answered in polynomial time in data complexity [7]. Decidability is preserved since only the certain answers are propagated through conjunctive peer mappings (see [7] for more details). We argue that this is an advantage of our approach, since an arbitrary interconnection of conjunctive peer mappings leads to undecidability if interpreted in FO semantics. Although they preserve decidability, we show that our peer mappings define non-FO-rewritable constraints, and so it is not possible to process queries in general RPSs by rewriting them into SPARQL queries. In this regard, for the preliminary empirical evaluation, we implement a SPARQL rewriting algorithm that leverages only FO-rewritable sets of peer mappings. As future work, we aim to investigate rewriting algorithms that produce queries in a language more expressive than FO-queries, in order to implement the full semantics of our system.

3 Implementation of the Proposed Approach

This section illustrates the main components of a middleware for LOD integration, which is also proposed in our recent paper [6]. The system provides a query inter-

[1] http://sameas.org

face between the user and the Linked Data sources and it is based on the formalisation of the RPS illustrated in the previous section. To summarise, our middleware exposes a unified view of heterogeneous RDF sources which are semantically linked with the RPS mapping assertions. A unified SPARQL endpoint accepts queries expressed in any source vocabulary. A SPARQL query rewriting engine rewrites the queries with respect to the semantic mappings of an instance of RPS, so as to retrieve more complete answers. Then, the rewritten query is evaluated over the sources in a federated approach and the query result is presented to the user.

In our system, the query rewriting engine is composed of two sub-engines. *(i)* The *semantic integration* module generates a "perfect rewriting" of the user's query, that is, a query that returns, once evaluated, a sound and complete answer of the original query based on the semantic mappings in the RPS. *(ii)* The *query federation* module executes a second rewriting step adopting the SPARQL 1.1 extension and exploiting the `SERVICE` clause; it generates a federated query to be evaluated over multiple RDF sources.

The system provides for *automated alignment* of the peer schemas, to link entities and concepts in the Linked Open Data cloud. It extracts structural information from the sources, such as the sets of entities, predicates, classes etc. Then, it performs schema alignment and coreference resolution by: *(i)* retrieving mappings between sources, such as `owl:sameAs` or `VoID`[2] triples, and other semantic links between sources; *(ii)* generating new mappings, using existing ontology matching and instance linkage techniques, such as *Falcon-AO* [11]; *(iii)* translating these alignments into our peer mapping language; and *(iv)* storing the mappings in the RPS.

For a preliminary empirical evaluation, we implemented a partial version of system that leverages only FO-rewritable sets of peer mappings which are manually designed. For this setting, we develop a SPARQL rewriting algorithm, called RPS-rewrite, which is based on a backward chaining mechanism [8]. It takes as input a conjunctive SPARQL query and it generates the FO-rewriting of the input query with respect to an instance of RPS, as a union of conjunctive SPARQL queries. Furthermore, we address two query optimisations of the query resulting from RPS-rewrite. The first optimisation performs a pruning of all the SPARQL disjuncts with triple patterns that cannot provide a successful graph pattern match. The second optimisation is given by "ignoring" the equivalence mappings during the backward chaining steps, since they lead to a production of SPARQL disjuncts that grow exponentially with respect to the number of mapping assertions. Consequently, equivalence mappings are treated as stored `sameAs` triples and leveraged on query evaluation for co-reference resolution, by adopting a technique of variable rewriting which we omit for space reasons. These `sameAs` triples are stored externally on a *Virtuoso* server and are accessed through a SPARQL endpoint.

Regarding query federation, triple patterns in the body of the query are then grouped with respect to the RDF sources that can provide a successful graph pattern match. Then, the groups are assigned to the endpoints of the related sources, and evaluated using the SPARQL 1.1 `SERVICE` clause. Finally, the results are presented to the user.

[2] http://www.w3.org/TR/void/

4 Empirical Evaluation Methodology

The goal of the preliminary evaluation is to provide a study of the behaviour of the current version of the framework with the aim of *(i)* ensuring that our framework can be used in its restricted version and, *(ii)* analysing basic performance in terms of cost execution time, and *(iii)* detecting current weaknesses of our framework to suggest future developments. We select three large-scale datastores with overlapping vocabularies in the domain of movies: *DBpedia*, *Linked Movie Database* and *Fact Forge*. The current version of our middleware is a *Java* application that takes as input a SPARQL query, generates the rewriting in SPARQL 1.1 and executes the rewritten federated query over the selected datastores using *Apache Jena*, the well-known open source Semantic Web framework for Java.

We performed a partial semantic alignment of DBpedia, Linked Movie Database and Fact Forge schemas, defining a set of FO-rewritable one-to-one mappings for similar classes and predicates, adopting the RPS mapping language. For instance, we define a mapping of the form:
$q(x, y) \leftarrow (x, \texttt{linkedmdb:actor}, y) \rightsquigarrow q(x, y) \leftarrow (x, \texttt{dbpedia:starring}, y)$
to express a 1:1 *predicate mapping* from the IRI `linkedmdb:actor` to the IRI `dbpedia:starring` , and, a mapping of the form:
$q(x) \leftarrow (x, \texttt{rdf:type}, \texttt{ff:Person}) \rightsquigarrow q(x) \leftarrow (x, \texttt{rdf:type}, \texttt{foaf:Person})$
to express a 1:1 *class mapping* from the IRI `ff:Person` to the IRI `foaf:Person`, leveraging the semantics of the built-in RDF predicate `rdf:type` for the class mappings. Also, we retrieved some `sameAs` triples from the sources and we generated new triples so as to encode the reflexive, symmetric and transitive closure of the `sameAs` binary relation; this provides co-reference resolution of IRIs as we explained in the previous section. The peer mappings obtained present arbitrary topologies and include some mapping cycles.

To conduct our tests, we generate a set of SPARQL queries with up to three triple patterns in the body. We then evaluate the queries over the three endpoints of the datastores in order to obtain our baselines. Finally we execute the queries on our middleware, and we compare the number of results retrieved and the query execution time. The results are shown in Figure 1 and allow us to derive two main insights. As expected, the amount of information retrieved increases significantly by adopting our system, due to its interoperability with heterogeneous vocabularies. In addition, the approach does not compromise query execution time, since overall the response time of our system can be seen as an average of the query response time over the single datastores. In fact, using the RPS can sometimes be faster than using just one single source endpoint. This may be due to the minimisation of the number of distributed-joins performed by Jena, which may increase the throughput with respect to the fastest endpoint (in our case DBpedia).

5 Lessons Learned, Open Issues, and Future Directions

In this paper we address the problem of integrating RDF data sources in a peer-based fashion, where mappings are defined between arbitrary peers, without a centralised

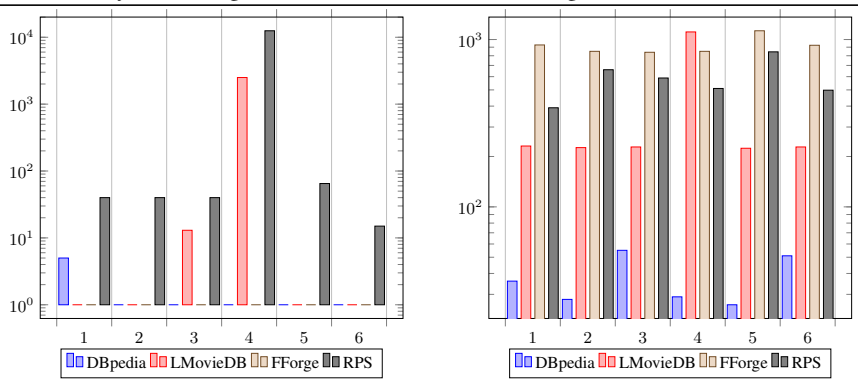

Fig. 1. Number of results on the left and query execution time on the right (logarithmic scales). Queries 1 - 6 shown on the x axes.

schema. We have proposed a novel formalisation of the notion of a peer-to-peer semantic integration system based on the Linked Open Data scenario. Following that, we have shown that query answering can be done in polynomial time in data complexity, and so we address Research Question RQ1 of this proposal. Finally, we have seen that it is not possible to process queries in general RDF peer systems by rewriting them into FO-queries, so we have conducted a preliminary empirical evaluation on FO-rewritable ontologies.

To address Research Question RQ2, we plan to devise a query rewriting algorithm that exploits the full semantics of our system. Firstly, we intend to investigate the possibility of adopting a *combined* approach, where the sources are partially materialised and queries are rewritten according to some of the dependencies only. To compute the perfect rewriting, another possible approach is to devise a rewriting algorithm that produces rewritten queries in a language more expressive than FO-queries, for instance Datalog, similarly to the approach in [4] which leverages new semantics for peer-to-peer systems based on epistemic logic. Another possible target language for the rewriting algorithm is SPARQL 1.1 with property paths; the idea is to leverage the expressive power of regular path queries in order to catch non-FO-rewritable constraints, such as the transitive closure of a relations. Two hypotheses follow from this approach: *(a)* it is possible to generate a SPARQL 1.1 query as a perfect rewriting of a conjunctive SPARQL query with respect to an RPS; *(b)* SPARQL 1.1 is not expressive enough: in this case we aim to characterise subsets of RPS mappings that are rewritable in SPARQL 1.1 with property paths.

Following from this, we will address Research Question RQ3. If the target language is SPARQL 1.1, query evaluation over multiple SPARQL endpoints can be done straightforwardly by exploiting the SPARQL 1.1 `SERVICE` clause. For Datalog rewritings, we will tackle the problem of distributed Datalog query processing on top of SPARQL endpoints.

Acknowledgments. I wish to thank my supervisors, Dr. Andrea Calì, Prof. Alexandra Poulovassilis and Dr. Peter Wood, for their invaluable support. I am also grateful to the reviewers for their constructive feedback.

References

1. Abiteboul, S., Duschka, O.M.: Complexity of answering queries using materialized views. In: Proc. of PODS. pp. 254–263 (1998)
2. Cai, M., Frank, M.: RDFPeers: a scalable distributed RDF repository based on a structured peer-to-peer network. In: Proc. of WWW. pp. 650–657 (2004)
3. Cai, M., Frank, M., Yan, B., MacGregor, R.: A subscribable peer-to-peer RDF repository for distributed metadata management. Web Semantics 2(2), 109–130 (2004)
4. Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Logical foundations of peer-to-peer data integration. In: Proc. of PODS. pp. 241–251 (2004)
5. Correndo, G., Salvadores, M., Millard, I., Glaser, H., Shadbolt, N.: SPARQL query rewriting for implementing data integration over Linked Data. In: Proc. of EDBT/ICDT Wksp (2010)
6. Dimartino, M.M., Calì, A., Poulovassilis, A., Wood, P.T.: Implementing peer-to-peer semantic integration of Linked Data. In: Proc. of BICOD (2015)
7. Dimartino, M.M., Calì, A., Poulovassilis, A., Wood, P.T.: Peer-to-peer semantic integration of Linked Data. In: Proc. of EDBT/ICDT Workshops. pp. 213–220 (2015)
8. Gottlob, G., Orsi, G., Pieris, A.: Ontological queries: Rewriting and optimization. In: Proc. of ICDE. pp. 2–13 (2011)
9. Halpin, H.: Identity, reference, and meaning on the web. In: Proc. of WWW Workshops (2006)
10. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology 1(1), 1–136 (2011)
11. Hu, W., Qu, Y., Cheng, G.: Matching large ontologies: A divide-and-conquer approach. Data & Knowledge Engineering 67(1), 140–160 (2008)
12. Le, W., Duan, S., Kementsietsidis, A., Li, F., Wang, M.: Rewriting queries on SPARQL views. In: Proc. of WWW. pp. 655–664 (2011)
13. Lenzerini, M.: Data integration: A theoretical perspective. In: Proc. of PODS. pp. 233–246 (2002)
14. Lopes, F.L.R., Sacramento, E.R., Lóscio, B.F.: Using heterogeneous mappings for rewriting SPARQL queries. In: DEXA Workshops. pp. 267–271 (2012)
15. Makris, K., Bikakis, N., Gioldasis, N., Christodoulakis, S.: SPARQL-RW: transparent query access over mapped RDF data sources. In: Proc. of EDBT. pp. 610–613 (2012)
16. Makris, K., Gioldasis, N., Bikakis, N., Christodoulakis, S.: Ontology mapping and SPARQL rewriting for querying federated RDF data sources. OTM pp. 1108–1117 (2010)
17. Montoya, G., Ibáñez, L.D., Skaf-Molli, H., Molli, P., Vidal, M.E.: SemLAV: local-as-view mediation for SPARQL queries. TLDKS Journal XIII pp. 33–58 (2014)
18. Nejdl, W., Wolpers, M., Siberski, W., Schmitz, C., Schlosser, M., Brunkhorst, I., Löser, A.: Super-peer-based routing strategies for RDF-based peer-to-peer networks. Web Semantics: Science, Services and Agents on the World Wide Web 1(2), 177–186 (2004)
19. Schenner, G., Bischof, S., Polleres, A., Steyskal, S.: Integrating distributed configurations with RDFS and SPARQL. In: 16th International Configuration Workshop. p. 9 (2014)
20. Schlegel, K., Stegmaier, F., Bayerl, S., Granitzer, M., Kosch, H.: Balloon fusion: SPARQL rewriting based on unified co-reference information. In: Proc. of ICDEW Workshops (2014)
21. Torre-Bastida, A.I., Bermúdez, J., Illarramendi, A., Mena, E., González, M.: Query rewriting for an incremental search in heterogeneous Linked Data sources. Flexible Query Answering Systems pp. 13–24 (2013)

Improving discovery in Life Sciences Linked Open Data Cloud

Ali Hasnain[1]

Insight Center for Data Analytics, National University of Ireland, Galway
firstname.lastname@insight-centre.org

Abstract. Multiple datasets that add high value to biomedical research have been exposed on the web as part of the Life Sciences Linked Open Data (LSLOD) Cloud. The ability to easily navigate through these datasets is crucial for personalized medicine and the improvement of drug discovery process. However, navigating these multiple datasets is not trivial as most of these are only available as isolated SPARQL endpoints with very little vocabulary reuse. The content that is indexed through these endpoints is scarce, making the indexed dataset opaque for users. We propose an approach to create an active Linked Life Sciences Data Compendium, a set of configurable rules which can be used to discover links between biological entities in the LSLOD cloud. We have catalogued and linked concepts and properties from 137 public SPARQL endpoints. Our Compendium is primarily used to dynamically assemble queries retrieving data from multiple SPARQL endpoints simultaneously.

1 Scene Setting

A considerable portion of the Linked Open Data cloud is comprised of datasets from Life Sciences Linked Open Data (LSLOD). The significant contributors includes the Bio2RDF project[1], Linked Life Data[2] and the W3C HCLSIG Linking Open Drug Data (LODD) effort[3]. The deluge of biomedical data in the last few years, partially due to the advent of high-throughput gene sequencing technologies, has been a primary motivation for these efforts. There had been a critical requirement for a single interface, either programmatic or otherwise, to access the Life Sciences (LS) data. Although publishing datasets as RDF is a necessary step towards unified querying of biological datasets, it is not sufficient to retrieve meaningful information due to data being heterogeneously available at different endpoints [2]. Moreover in the LS domain, LD is extremely heterogeneous and dynamic [14,6]; also there is a recurrent need for *ad hoc* integration of novel experimental datasets due to the speed at which technologies for data capturing in this domain are evolving. As such, integrative solutions increasingly rely on federation of queries [4]. With the standardization of SPARQL 1.1, it is now possible to assemble federated queries using the "SERVICE" keyword, already supported by multiple tool-sets (SWobjects and Fuseki etc). To assemble queries encompassing multiple graphs distributed over different places, it is necessary

[1] http://bio2rdf.org/ (l.a.: 2015-03-31)
[2] http://linkedlifedata.com/ (l.a.: 2015-05-16)
[3] http://www.w3.org/wiki/HCLSIG/LODD (l.a.: 2014-07-16)

that all datasets should be query-able using the same global schema [15]. This can be achieved either by ensuring that the multiple datasets make use of the same vocabularies and ontologies, an approach previously described as *"a priori integration"* or conversely, using *"a posteriori integration"*, which makes use of mapping rules that change the topology of remote graphs to match the global schema [5]. The methodology to facilitate the latter approach is the focus of our research. Moreover for LD to become a core technology in the LS domain, three challenges need to be addressed: *i)* dynamically discover datasets containing data regarding biological entities (e.g. Drugs, Molecules), *ii)* retrieve information about the same entities from multiple sources using different schemas, and *iii)* identify, for a given query, the highest quality data.

To address the aforementioned challenges, we introduce the notion of an active Compendium for LS data – a representation of entities and the links connecting these. Our methodology consisted of two steps: *i)* catalogue development, in which metadata is collected and analyzed, and *ii)* links creation, which ensures that concepts and properties are properly mapped to a set of Query Elements (Qe) [17]. For evaluation purposes, Qe are defined in the context of Drug Discovery and can be replaced by other *Qe(s)*. We assume that the proposed Compendium holds the potential to be used for a number of practical applications including assembling federated queries in a particular context. We already proposed the Link Creation mechanism, approaches and the linking statistics [7] as well as the cataloguing mechanism [9] and in this article we briefly report the methodology, initial results for Compendium development (cataloguing and linking), and an architecture for implementing Domain Specific Query Engine (under progress) as one of the practical applications that federates SPARQL query based on the set of mapping rules defined in the Compendium.

2 State of the Art

One approach to facilitate the *"A posteriori integration"* is through the use of available schema: semantic information systems have used ontologies to represent domain-specific knowledge and enable users to select ontology terms in query assembly [11]. BLOOMS, for example, finds schema-level links between LOD datasets using ontology alignment [10], but it relies mainly on Wikipedia. Ontology alignment typically relies on starting with a single ontology, which is not available for most SPARQL endpoints in the LOD cloud, hence could not be applied in our case. Furthermore, ontology alignment does not make use of domain rules (e.g. for two same sequences, qualifies for same gene) nor the use of URI pattern matching for alignment [7]. Approaches such as the VoID [1] and the SILK Framework [16] enable the identification of rules for link creation, but require extensive knowledge of the data prior to links creation. Query federation approaches have developed some techniques to meet the requirements of efficient query computation in the distributed environment. FedX [13], a project which extends the Sesame Framework [3] with a federation layer, enables efficient query processing on distributed LOD sources by relying on the assembly of a catalogue of SPARQL endpoints but does not use domain rules for links creation. Our approach for link creation towards Compendium development is a combination of

the several linking approaches as already explained by Hasnain et. al [7]: *i)* similarly to ontology alignment, we make use of label matching to discover concepts in LOD that should be mapped to a set of *Qe*, *ii)* we create *"bags of words"* for discovery of schema-level links similar to the approach taken by BLOOMS, and *iii)* as in SILK, we create domain rules that enable the discovery of links.

3 Proposed Approach/ Methodology

We proposed a Compendium for navigating the LSLOD cloud. Our methodology consists of two stages namely catalogue generation and link generation. Data was retrieved from 137 public SPARQL endpoints[4] and organized in an RDF document - the LSLOD Catalogue. The list of SPARQL endpoints was captured from publicly available Bio2RDF datasets and Datahub[5].

3.1 Methodology for Catalogue Development

For cataloguing, a preliminary analysis of multiple public SPARQL Endpoints was undertaken and a semi-automated method was devised to retrieve all classes (concepts) and associated properties (attributes) available by probing data instances. The workflow definition is as follows:

1. For every SPARQL endpoint S_i, find the distinct Classes $C(S_i)$:

$$C(S_i) = Distinct\ (Project\ (?class\ (toList\ (BGP\ (triple\ [\ ]\ a\ ?class\ )))))) \quad (1)$$

2. Collect the Instances for each Class $C_j(S_i)$:

$$I_i : C_j(S_i) = Slice\ (Project\ (?I\ (toList\ (BGP\ (triple\ ?a\ a\ < C_j(S_i) >)))),\ rand()) \quad (2)$$

3. Retrieve the Predicate/Objects pairs for each $Ii : C_j(S_i)$:

$$I_i(P,O) = Distinct\ (Project\ (?p,\ ?o\ (toList\ (BGP\ (triple\ < I_i : C_j(S_i) > ?p\ ?o\ )))) \quad (3)$$

4. Assign Class $C_j(S_i)$ as domain of the Property P_k :

$$Domain(P_k) = C_j(S_i) \quad (4)$$

5. Retrieve Object type (O_T) and assign as a range of the Property P_k :

$$Range(P_k) = O_T; O_T = \begin{cases} rdf : Literal & \text{if } (O_k\ is\ String) \\ dc : Image & \text{if } (O_k\ is\ Image) \\ dc : InteractiveResource & \text{if } (O_k\ is\ URL) \\ Project\ (?R\ (toList\ (BGP & \\ (triple\ < O_k >\ rdf : type\ ?R))) & \text{if } (O_k\ is\ IRI) \end{cases} \quad (5)$$

RDFS, Dublin Core[6] and VoID[7] vocabularies were used for representing the data in the LSLOD catalogue, a slice of the catalogue for Pubchem SPARQL endpoint is presented (fig.1).

[4] `http://goo.gl/ZLbLzq`
[5] `http://datahub.io/` (l.a.: 2015-05-05)
[6] `http://dublincore.org/documents/dcmi-terms/` (l.a.: 2014-07-12)
[7] `http://vocab.deri.ie/void` (l.a.: 2014-07-12)

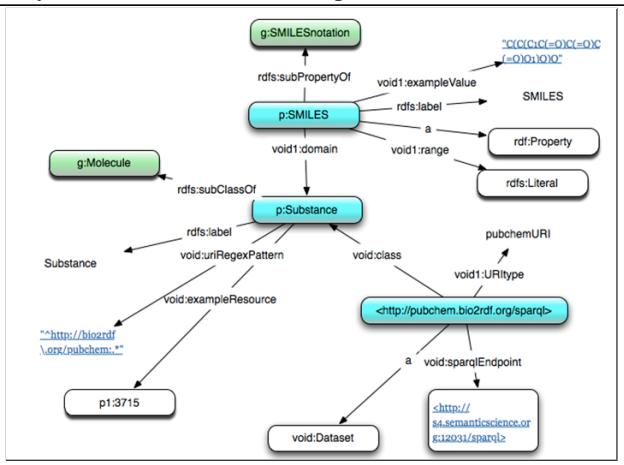

Fig. 1: An Extract from the LSLOD Catalogue for Pubchem dataset

3.2 Methodology for Link Generation

During this phase `subClassOf` and `subPropertyOf` links were created amongst different concepts and properties to facilitate *"a posteriori integration"*. The creation of links between identified entities (both chemical and biological) is not only useful for entity identification, but also for discovery of new associations such as protein/drug, drug/drug or protein/protein interactions that may not be obvious by analyzing datasets individually. Figure. 1 shows the `subClassOf` and `subPropertyOf` links with defined *Qes*. Links were created (discussed previously in [7]) using several approaches: *i)* Naïve Matching/ Syntactic Matching/ Label Matching, *ii)* Named Entity Matching, *iii)* Domain dependent/ unique identifier Matching, and *iv)* Regex Matching.

4 Applications/ Current Implementation

As of 31^{st} May 2015, the Compendium consists of 280064 triples representing 1861 distinct classes and 3299 distinct properties catalogued from 137 endpoints.

4.1 DSQE: Domain-specific Query Engine

The general architecture of the DSQE (Fig 2) shows that given a SPARQL query, the first step is to parse the query and get individual triple patterns. Then Compendium is used for the triple pattern wise source selection (TPWSS) to identify relevant sources against individual triple patterns of the query. The Compendium enumerates the known endpoints relates each endpoint with one or more graphs and maps the local vocabulary to the vocabulary of the graph. The resulting query is executed on top of Apache Jena query engine.

An instance[8] of the DSQE is deployed in the context of drug discovery [8]. Using 'Standard' query builder, the user can select a topic of interest (e.g.

[8] `http://srvgal86.deri.ie:8000/graph/Granatum`

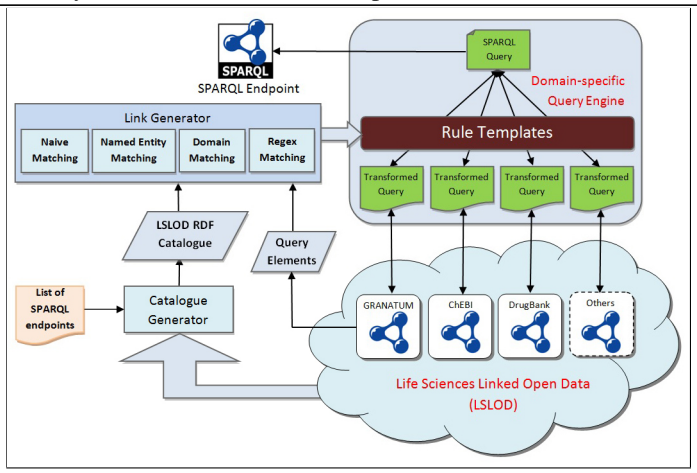

Fig. 2: Compendium and Query Engine Architecture

`Molecule`) along with the list of associated properties. We plot the catalogued subclasses of few Qe and the total number of distinct instances retrieved per Qe while querying using DSQE (Fig 3).

5 Evaluation

So far we evaluated the performance of our catalogue generation methodology and recorded the times taken to probe instances through endpoint analysis of 12 endpoints whose underlying data sources were considered relevant for drug discovery - Medicare, Dailymed, Diseasome, DrugBank, LinkedCT, Sider, National Drug Code Directory (NDC), SABIO-RK, Saccharomyces Genome Database (SGD), KEGG, ChEBI and Affymetrix probesets. The cataloguing experiments were carried out on a standard machine with 1.60Ghz processor, 8GB RAM using a 10Mbps internet connection. We recorded the total available concepts and properties at each SPARQL endpoint as well as those actually catalogued in our Compendium [9]. Total number of triples exposed at each of these SPARQL endpoints and the time taken for cataloguing was also recorded. We selected those SPARQL endpoints which have a better latency for this evaluation, as the availability and the uptime of the SPARQL endpoint is an important factor for cataloguing. Best fit regression models were then calculated. As shown in Fig. 4, our methodology took less than 1000000 milliseconds (<16 minutes) to catalogue seven of the SPARQL endpoints, and a gradual rise with the increase in the number of available concepts and properties. We obtained two power regression models ($T = 29206 * C_n^{1.113}$ and $T = 7930 * P_n^{1.027}$) to help extrapolate time taken to catalogue any SPARQL endpoint with a fixed set of available concepts (C_n) and properties (P_n), with R^2 values of 0.641 and 0.547 respectively. Using these models and knowing the total number of available concepts/properties, a developer could determine the approximate time (ms) as a vector combination. KEGG and SGD endpoints took an abnormally large amount of time for cataloguing than the trendline. The reason for this may include endpoint timeouts or network

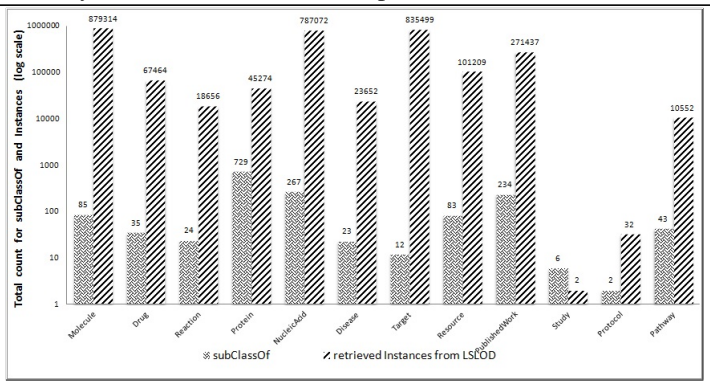

Fig. 3: Number of retrieved Instances and Subclasses linked to any *Qe*

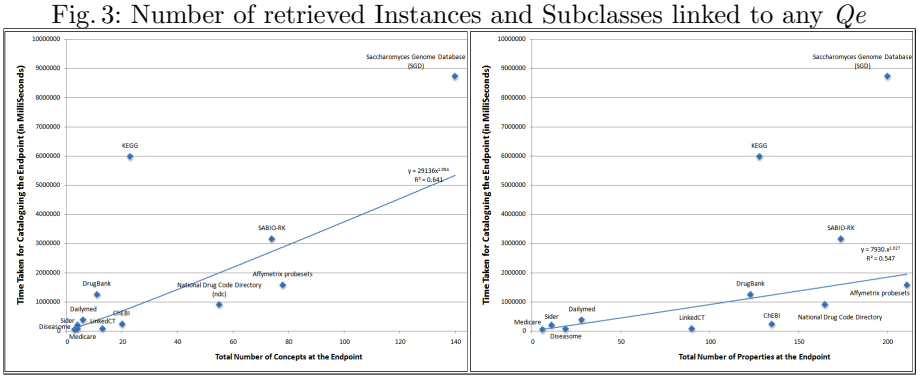

Fig. 4: Time taken to catalogue 12 SPARQL endpoints

delays. We also evaluated the performance of our Link Generation methodology by comparing it against the popular linking approaches. Using WordNet thesauri we attempted to automate the creation of bags of related words using 6 algorithms [7]: Jing Conrath, Lin, Path, Resnik, Vector and WuPalmer with unsatisfactory results (Figure 5(c)). Our linking approaches resulted in better linking rate as shown in Figure 5(a,b)

6 Discussion

There is great potential in using semantic web and LD technologies for accessing and querying Life sciences data for finding Meaningful Biological Correlations. However, in most cases, it is not possible to predict *a priori* where the relevant data is available and its representation. Our current research provides the concept and methodology for devising an active Linked Life Sciences Data Compendium that relies on systematically issuing queries on various life sciences SPARQL endpoints and collecting its results in an approach that would otherwise have to be encoded manually by domain experts. Current experiments and evaluation uses a set of *Qe*, which were defined in a context of drug discovery. The number of classes per endpoint varied from a single class to a few thousands. Our initial exploration of the LSLOD revealed that only 15% of classes are reused. However, this was not the case for properties, of which 48.5% are reused. Most

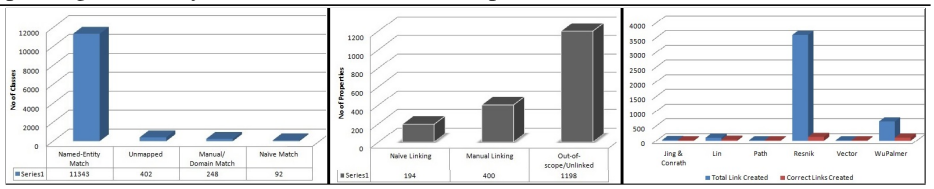

Fig. 5: (a) Number of Classes Linked, (b) Number of Properties Linked, (c) Number of Classes linked through available similarity linking approaches

of the properties found were domain independent (e.g. `type`, `seeAlso`); however, these are not relevant for the Compendium as they cannot increase the richness of information content. Although a very low percentage of linking becomes possible through naïve matching or manual/domain matching, the quality of links created are highly trusted [7]. It is also worth noticing that 23% of identified classes, and 56.2% of the properties remained unlinked, either because they are out of scope or cannot match any *Qe*. This means that the quality as well as the quantity of links created is highly dependent on the set of *Qe* used.

7 Open Issues and Future Directions

Multiple challenges faced which can hinder the applicability of our approach:

- Some endpoints return timeout errors when a simple query (`SELECT DISTINCT ?Concept WHERE {[ ] a ?Concept}`) is issued.
- Some endpoints have high downtime and cannot be generally relied.
- Many endpoints provide non-deferenceable URI and some derefenceable URI do not provide a "type" for the instance.

In future an extension under consideration to available Compendium is to enrich it with statistical and provenance information with appropriate changes to DSQE and evaluate the overall performance. This includes information including *void:triples*, *void:entities*, *void:classes*, *void:properties*, *void:distinctSubjects* and *void:distinctObjects* in case of statistical cataloguing where as *dcterms:title*, *dcterms:description*, *dcterms:date*, *dcterms:publisher*, *dcterms:contributer*, *dcterms:source*, *dcterms:creator*,
dcterms:created, *dcterms:issued* and *dcterms:modified* in case of provenance. Currently we are extending DSQE to convert any SPARQL 1.0 query into corresponding SPARQL 1.1 query by using TPWSS information and the SPARQL "SERVICE" clause. Implementing so DSQE will be able to answer any federated SPARQL Query considering the desired endpoint being catalogued in Compendium. The performance of this extended DSQE is aimed to compare with state of the art query Engine FedX [13] using extensive evaluation criteria including *source selection in terms of number of ASK, total triple pattern-wise sources selected, source selection time and total number of results retrieved per query*. For this evaluation we aim to select some queries from available query federation benchmark e.g FedBench [12] and also plan to define some complex biological queries applicable on 10 real time publicly available datasets. Issues related to Identity Resolution are also considered as future work.

Acknowledgement

This research has been supported in part by Science Foundation Ireland under Grant Number SFI/12/RC/2289 and SFI/08/CE/I1380 (Lion 2). The author would also like to acknowledge Stefan Decker being PhD supervisor.

References

1. Alexander, K., Hausenblas, M.: Describing linked datasets-on the design and usage of void, the'vocabulary of interlinked datasets. In: In Linked Data on the Web Workshop (LDOW 09), in conjunction with WWW09. Citeseer (2009)
2. Bechhofer, S., Buchan, I., De Roure, D., Missier, P., et al.: Why linked data is not enough for scientists. Future Generation Computer Systems 29(2), 599–611 (2013)
3. Broekstra, J., Kampman, A., Van Harmelen, F.: Sesame: A generic architecture for storing and querying RDF and RDF schema. In: The Semantic Web—ISWC 2002, pp. 54–68. Springer (2002)
4. Cheung, K.H., Frost, H.R., Marshall, M.S., et al.: A journey to semantic web query federation in the life sciences. BMC bioinformatics 10(Suppl 10), S10 (2009)
5. Deus, H.F., Prud'hommeaux, E., Miller, M., Zhao, J., Malone, J., Adamusiak, T., et al.: Translating standards into practice–one semantic web API for gene expression. Journal of biomedical informatics 45(4), 782–794 (2012)
6. Goble, C., Stevens, R., Hull, D., et al.: Data curation+ process curation= data integration+ science. Briefings in bioinformatics 9(6), 506–517 (2008)
7. Hasnain, A., Fox, R., Decker, S., Deus, H.F.: Cataloguing and linking life sciences LOD Cloud. In: 1st International Workshop on Ontology Engineering in a Data-driven World collocated with EKAW12 (2012)
8. Hasnain, A., Kamdar, M.R., Hasapis, P., Zeginis, D., Warren Jr, C.N., et al.: Linked Biomedical Dataspace: Lessons Learned integrating Data for Drug Discovery. In: International Semantic Web Conference (In-Use Track), October 2014 (2014)
9. Hasnain, A., e Zainab, S.S., Kamdar, M.R., Mehmood, Q., Warren Jr, C.N., Fatimah, Q.A., Deus, H.F., Mehdi, M., Decker, S.: A roadmap for navigating the life sciences linked open data cloud. In: Semantic Technology, pp. 97–112. Springer (2014)
10. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology alignment for linked open data. In: The Semantic Web–ISWC 2010, pp. 402–417. Springer (2010)
11. Petrovic, M., Burcea, I., Jacobsen, H.A.: S-ToPSS: semantic toronto publish/subscribe system. In: Proceedings of the 29th international conference on Very large data bases-Volume 29. pp. 1101–1104. VLDB Endowment (2003)
12. Schmidt, M., Görlitz, O., Haase, P., Ladwig, G., Schwarte, A., Tran, T.: Fedbench: A benchmark suite for federated semantic data query processing. In: The Semantic Web–ISWC 2011, pp. 585–600. Springer (2011)
13. Schwarte, A., Haase, P., Hose, K., Schenkel, R., Schmidt, M.: Fedx: a federation layer for distributed query processing on linked open data. In: The Semanic Web: Research and Applications, pp. 481–486. Springer (2011)
14. Stein, L.D.: Integrating biological databases. Nature Reviews Genetics 4(5), 337–345 (2003)
15. Studer, R., Grimm, S., Abecker, A.: Semantic web services: concepts, technologies, and applications. Springer (2007)
16. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and maintaining links on the web of data. Springer (2009)
17. Zeginis, D., et al.: A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web (2013)

Entity Linking and Knowledge Discovery in Microblogs

Pikakshi Manchanda

Department of Computer Science, Systems and Communication,
Università di Milano-Bicocca, Milano, Italy
pikakshi.manchanda@disco.unimib.it

Abstract. Social media platforms have become significantly popular and are widely used for various customer services and communication. As a result, they experience a real-time emergence of new entities, ranging from product launches to trending mentions of celebrities. On the other hand, a Knowledge Base (KB) is used to represent entities of interest/relevance for general public, however, unlikely to cover all entities appearing on social media. One of the key tasks towards bridging the gap between Web of Unstructured Data and Web of Data is identifying such entities from social media streams which are important and haven't been yet represented in a KB. The main focus of this PhD work is discovery of new knowledge from social media streams in the form of new entities and/or new mentions of existing entities while enriching KBs as well as lexically extending them for existing entities. Based on the discovery of new entities or new mentions, structured data in the form of RDF (Resource Description Framework) can be extracted from the Web.

Key words: Social Media, Knowledge Base, Web of Data, RDF

1 Scene Setting

1.1 Objective of the Research

Microblogging platforms have become an indispensable resource for users by providing services such as sales and marketing, news and communication, trend detection and a variety of customer services. Due to their dynamic nature, they experience a steady emergence of new knowledge in the form of new entities (such as product launches), new relations between existing entities (such as a football player playing for *FC Barcelona* and *Real Madrid*), as well as new/popular mentions of existing entities (such as trending colloquial names for celebrities). Knowledge bases provide a broad (yet intrinsically non exhaustive) coverage of a variety of entities found on the Web and social media streams. However, it is unlikely that a KB can provide coverage of all new entities that emerge constantly on the Web. As a result, tasks such as Named Entity Recognition (NER), Disambiguation (NED) and Linking (NEL) have gained significant attention of NLP practitioners. Named entity recognition is the task of identifying a piece of

text as a named entity and classifying into types such as person, location, organization etc. whereas a named entity disambiguation task is to disambiguate a named entity with a resource in a KB and finally link it with the said resource.

In order to enrich a KB for new/relevant entities emerging on social media in real-time, it is necessary to identify those entities and gather contextual information from the Web and social media. The objective of this work is not only to identify and extract new knowledge, but also being able to use it in order to enrich and lexically extend KBs. In the process, we will be able to improve the accuracy of named entity recognition as well as disambiguation tasks.

1.2 Research Questions

The proposed research work aims to address the following research questions:
RQ1: Can we perform NER and NEL in microposts as a joint task and link the named entities to resources in a KB?
RQ2: Is it possible to use the results of an Information Extraction (IE) task to identify new entities?
RQ3: Can we use an enriched/lexically extended KB to improve the IE process of new entities from microblogging platforms?

1.3 Motivation and Relevance

Significant gain in momentum for IE (achieved mainly through NER and NEL), from news archives, blogs and Web pages, is attributed to need for bridging the gap between Document Web and Web of Data. The main motivation for carrying out a research on discovery of new knowledge by means of IE tasks is primarily because new entities emerge frequently over social media. Another motivating factor is being able to perform entity recognition and disambiguation on short textual formats, such as microblogs, as a joint task in an end-to-end entity linking pipeline. This is also important from the point of view of KB enrichment and its lexical extension for existing entities.

KB Enrichment can be performed automatically to some extent (by identifying a new entity, and collecting contextual information from the Web) or can even be performed interactively, for instance, driven by social content creation communities. The output of my research work combined with these techniques can be used to enrich KBs periodically. Furthermore, a lot of research (Semantic Search, Recommendation Systems, Disaster Discovery, Sentiment Analysis) is dependent on entity disambiguation and discovery of new knowledge.

1.4 Challenges and Opportunities

Challenges: The task of identification and disambiguation of entities from microblogs is challenging due to the following reasons:

- *Short, noisy nature*: An informal microblogging style, coupled with use of Internet slang and misspellings [4, 7, 8] renders it difficult to identify new entities, affecting the *accuracy* of entity recognition and disambiguation.

- *Occurrence of Out Of Vocabulary (OOV) mentions*: We define an OOV mention as an existing resource in a KB, being referred by an alternate entity mention in social media which is not present in KB. As a result, OOV mentions can't be disambiguated, causing the performance accuracy of an end-to-end entity linking system to suffer.
- *Occurrence of Out of Knowledge base (OOKB) entities*: We define an OOKB entity as one which is not covered by a KB and, thus, can be considered as newly emerging.

Opportunities: If we are able to identify an OOV mention, we can *lexically enrich* the KB for said existing entity. Similarly, if we are able to detect an OOKB entity, we can *extensionally update* the KB for the new entity by collecting contextual information about it from the Web. On the other hand, by addressing the above challenges, we will also be able to improve the accuracy of the end-to-end entity linking pipeline.

2 Proposed Approach

2.1 Formal Definition and Properties of the Approach

Given a tweet t, the goal of the system is to identify named entities in t. Further, the system maps every identified entity e to a referent resource r in knowledge base K. More formally, we define a Named Entity Recognition task as a function which identifies and maps a set of words W in tweet t to a tuple of entity name, e_i, and a corresponding entity type, $type_{e_i}$, i.e.,

$$f_{NER} : W \rightarrow < e_i^t, type_{e_i}^t > \tag{1}$$

Next, we define a universe U consisting of entities present in unstructured/semi-structured data on social media and the Web as well as resources covered by KBs. Further, we define a Named Entity Linking task as a function which maps an identified entity e_i^t, as in equation (1), to a resource r_j in K, i.e.,

$$f_{NEL} : e_i^t \rightarrow r_j^K \tag{2}$$

Here every resource r_j in K can be associated with one or more resource types and is represented as c_{r_j}. f_{NEL} is defined for entities which are covered by resources in K. OOV mentions also have referent resources in K, however, the said mention has been referred in social media by an alternate name while K isn't lexically updated to provide coverage for it. It is to note here, that OOV mention, its original entity as well as the corresponding resource are present in U, however, f_{NEL} is unable to link the OOV mention to the corresponding resource. On the other hand, OOKB entities are new entities, present in U, which have not yet been covered by K and so f_{NEL} is unable to link them as well.

2.2 Relationship between your approach and state-of-art approaches

Various existing approaches [1, 5, 9] as well as a variety of commercial tools, such as Zemanta[1], Alchemy API[2], and DBpedia Spotlight[3] are used for entity recognition in text. However, these conventional tools perform poorly on short textual data [1], mainly due to lack of context and informal language. [8] propose a tweet-based NLP framework for entity recognition in tweets using a CRF model with the help of contextual, dictionary and orthographic features. In [5], Liu et al. (2011) propose an entity recognition framework using K-Nearest Neighbour (KNN) Classifier with a linear CRF Model.

State-of-the-art approaches provide a variety of methods for entity disambiguation. However, few existing approaches target the detection of new entities using existing knowledge provided by KBs. Liu et al (2013) use similarity measures to detect OOV mentions of existing entities [4], however, OOKB entities are not dealt with. [2] propose an end-to-end tweet-level identification and disambiguation system while using structural learning techniques to jointly optimize identification as well as disambiguation. However, their approach is not able to recognize or deal with OOKB entities. An approach for discovery of emerging OOKB entities with ambiguous names from documents has been proposed in [3]. This work is, in principle, a foundation for our research work, however, their approach doesn't consider the entities emerging in social media streams.

Based on the literature review, we observe that most state-of-the-art systems treat entity identification and disambiguation as separate tasks. In this research work, we propose an end-to-end entity linking pipeline where we study entity recognition and disambiguation as a joint problem for microposts. To the best of our knowledge, our work provides a novel contribution, in the sense, that we not only aim to improve the disambiguation of entities using linked datasets, but also we address the task of discovery of new entities from tweets, thus improving the overall accuracy of the system. We distinguish between OOV mentions and OOKB entities and also propose distinctive measures to deal with both types of entities. We use the discovered information for KB enrichment.

3 Implementation of the Proposed Approach

3.1 The Big Picture and Current Implementation

In this section, we present a brief overview of the system, as shown in Fig. 1. The system performs tweet-wise evaluation using Ritter et al's (2011) state-of-the-art T-NER system [8] for entity recognition and classification. Further, for NED, we have constructed an inverted index of the data property **rdfs:label** from DBpedia[4] which we currently consider for disambiguation and knowledge discovery.

[1] `http://www.zemanta.com/`
[2] `http://www.alchemyapi.com/`
[3] `http://dbpedia.org/spotlight/`
[4] `http://wiki.dbpedia.org/`

For every identified entity in a tweet, an ad-hoc index lookup is performed to obtain a list of candidate resources which we rank using a high-recall lookup approach. We use contextual and orthographic features, identified entity-type,

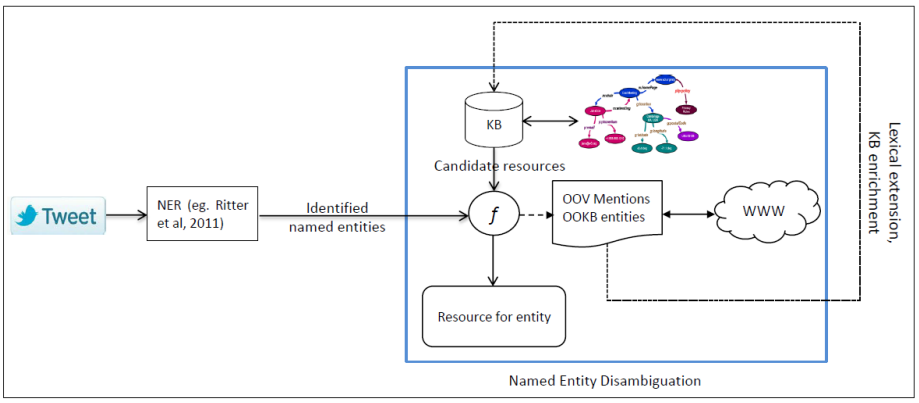

Fig. 1. Framework of the proposed model

as well as tweet-specific features such as use of @usernames, #hashtags and URLs which aid in disambiguation amongst the candidate resources. This is accomplished using a probabilistic matching function, being presented in this work, which takes into consideration the following factors:

1. Lexical Similarity between an entity in a tweet and candidate resource
2. Coherence between an entity and (structured) document page of candidate resource in KB
3. Relatedness between entities in a tweet, in case where there is more than one entity in a tweet

Currently, we have implemented measures to calculate similarity between entity in a tweet and candidate resources, as well as relatedness between entities (from KB perspective, i.e., how frequently entities mentioned in a tweet co-occur in a KB). In the future, we plan to take into consideration the relatedness between entities from real-world perspective.

The probabilistic matching function helps to disambiguate named entities with resources in KB. Subsequently, we will obtain a pool consisting of entities which can't be disambiguated. This pool will consist of OOV mentions, OOKB entities as well as noise (text wrongly identified as an entity). Information, such as usage patterns, frequency of usage, as well as contextual patterns from social media streams and the Web, will be collected for entities in pool. This information can be used to enrich a KB, either automatically or manually by content creation communities, with the help of which disambiguation is performed again to improve the overall accuracy of disambiguation process of the proposed system.

4 Empirical Evaluation Methodology

4.1 General Strategy

The research questions described above are related to a few hypotheses:
H1: If an entity is a word that appears in the lexicon of a resource, the system links it with the resource with a certain degree of accuracy. For this, we use entity information from tweet and resource in KB. In order to accomplish this task, we perform NER and NEL jointly (explained in detail in section 4.2). A NER system exhibits segmentation errors (such as *St. Mary's* identified as 2 distinct entities), identification erros (such as *justten* being identified as an entity) and classification errors (such as *Hawaii* being identified as Person). We use Ritter et al.'s (2011) gold standard corpus of 2400 tweets for NER. Additionally, we created a manually annotated gold standard collection of named entities for NEL from gold standard corpus used for NER.

H2: If there is a pool of unknown entities, we collect additional knowledge, from the Web and social media, in order to classify them as new (OOKB) entities or use that knowledge to resolve (OOV) entity mentions and link them with resources in KB. A gold standard corpus of such unknown entities needs to be created for this step. We can also use the pool of entities from NEL's gold standard which aren't disambiguated. We plan to expand this gold standard in the future.

4.2 Current State of the Evaluation

In this section, we present the evaluation results achieved so far for hypotheses H1. We plan to start creating a gold standard for H2 by December 2015.

H1-Task1: Entity Recognition (Experimental Analysis of T-NER)
Using Ritter et al.'s (2011) gold standard corpus of 2400 tweets, T-NER identifies a total of 1496 named entities classified into 10 distinct entity types (person, location, organization,..), in contrast to 1612 named entities as found in the ground truth. T-NER exhibits an identification error rate of 9.62%, whereas segmentation error rate is negligible. We summarize the classification error rate of every entity type in Table 1 below. As is evident, the classification error rate is quite high for entity types *Movie* and *Band*. A significant reason for this could be attributed to out-of-date knowledge utilized by T-NER for entity recognition.

H1-Task2: Entity Disambiguation (Experimental Analysis of Lexical Similarity Measure and Relatedness)

In this step, we use the set of named entities identified in Task 1 and based on an *ad-hoc candidate match retrieval approach*, we obtain candidate resources for these named entities from our index of rdfs:label. A manually annotated gold

Table 1. Classification Error rate for T-NER

Entity Type	Error (%)
Band	73.83
Company	21.9
Facility	54.79
Geo-Location	19.75
Movie	75.83
Other	46.29
Person	28.18
Product	39.70
Sportsteam	48.27
TVshow	48.71

Table 2. Precision-Recall of named entities with candidate matches

Entity Query Representation	P(%)	R%)
Entity Mention	92	95
Entity Mention and type	87	99
Combined Entity Mentions	31	24

standard of 1455 named entities is created out of 1496 entities that were identified in Task 1 to aid in candidate match retrieval. The remaining entities serve as a pool of unknown entities and need to be further expanded in order to be used as a gold standard for Task 3 described below. We have experimented with varying forms of entity representations (only entity mention, entity mention with its entity type, and a combination of entity mentions) in order to obtain sufficient number of candidate matches for each named entity. Table 2 summarizes the precision-recall for varying entity representations.

The first representation produces a list of candidate resources (highest precision) for disambiguation. Second representation produces a list with the highest recall (fetching noisy results as well), however, there is a decrease in precision. The reason for such an output can be due to knowledge gaps in KB for specific entity types (thus, a justified need for KB enrichment). The third representation is for tweets which have more than one entity. This representation exhibits the lowest precision as well as recall amongst all three. This can be due to infrequent occurrence of various entities together in social media, thus making it difficult to find sufficient evidences of their co-existence in a KB.

We implement a lexical similarity measure using *Lucene's Vector Space Model of Information Retrieval* to estimate similarity between an entity and a candidate resource so as to choose the most suitable resource for an entity. We have also used a relatedness measure in order to estimate co-occurrence frequency between two entities in a tweet, using a method described in [6]. Currently, we have implemented this measure from KB perspective, i.e., how often entities in a tweet can co-occur in a KB.

We found a total of 399 tweets in Ritter et al's dataset which have more than one entity. A high relatedness score depicts presence of a strong evidence in the KB that said entities co-occur frequently. Use of relatedness measure is attributed towards the need of improving the accuracy of disambiguation for infrequent/long-tail entities found in social media streams. Another significant reason for the use of this measure is in identifying an OOV entity mention.

H2-Task3: OOV Mention/OOKB Entity discovery
Discovery of OOV mentions as well as OOKB entities depends to a great extent on the performance accuracy of entity recognition as well as disambiguation. Herein, we propose to improve entity recognition by improving entity disambiguation, which is currently under progress. In order to achieve this, we use features (contextual information, evidences from KB, relatedness of an entity with other real-world entities) for entity recognition that are conventionally being used for entity disambiguation in the state-of-the-art. By improving entity recognition, the overall accuracy of the system will be improved.

5 Lessons Learned, Open Issues, and Future Directions

It is essential to discover new entities for the enrichment of KBs. However, one of the important lessons that we have learned is that, not every entity that has been discovered can be updated in a KB. Its authenticity needs to be verified as well as its relation and relevance w.r.t other entities in the real world has to be taken into consideration to update concepts in KBs.

Enrichment of KBs, specifically enriching the lexicon of an entity in a KB using information extracted from social media is one of the most important open issues in the Semantic Web community. As of now, we have conducted a variety of experiments for improving disambiguation. While we continue to improve it, the next step in this research work is working towards enrichment of KBs in time and extending them with quality information extracted from Social Media and the Web.

References

1. Derczynski, L., Maynard, D., Rizzo, G., van Erp, M., Gorrell, G., Troncy, R., Petrak, J., and Bontcheva, K. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* (2015).
2. Guo, S., Chang, M.-W., and Kiciman, E. To link or not to link? a study on end-to-end tweet entity linking. In *HLT-NAACL* (2013).
3. Hoffart, J., Altun, Y., and Weikum, G. Discovering emerging entities with ambiguous names. In *Proceedings of Conference on WWW* (2014).
4. Liu, X., Li, Y., Wu, H., Zhou, M., Wei, F., and Lu, Y. Entity linking for tweets. In *ACL (1)* (2013).
5. Liu, X., Zhang, S., Wei, F., and Zhou, M. Recognizing named entities in tweets. In *Proceedings of ACL: Human Language Technologies* (2011).
6. Medelyan, O., Witten, I. H., and Milne, D. Topic indexing with wikipedia. In *Proceedings of AAAI WikiAI workshop* (2008).
7. Meij, E., Weerkamp, W., and de Rijke, M. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining* (2012).
8. Ritter, A., Clark, S., Etzioni, O., et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on EMNLP* (2011).
9. Usbeck, R., Ngonga Ngomo, A.-C., Luo, W., and Wesemann, L. Multilingual disambiguation of named entities using linked data. In *International Semantic Web Conference* (2014).

Answering SPARQL Queries using Views

Gabriela Montoya

LINA– Université de Nantes, France
gabriela.montoya@univ-nantes.fr

Abstract. Views are used to optimize queries and to integrate data in Databases. The data integration schema is composed of terms, they are used to pose queries to the integration system, and to describe sources data. When the data descriptions are SPARQL conjunctive queries, their number and the complexity of answering queries using them may be very high. In order to keep query answering cost low, and increase the usability of views to answer queries, we make the assumption of the simplest form of replication among services, and use triple pattern views as the Linked Data Fragments and Col-graph approaches have recently done. We propose two approaches: the SemLAV approach to integrate heterogeneous data using Local-as-View mappings to describe Linked Data and Deep Web sources, and the FEDRA approach to process queries against federations of SPARQL endpoints with replicated fragments.

1 Scene setting

The Linked Data Cloud includes more than one thousand datasets.[1] However, a larger number of sources is available in the Web [14], and because of their heterogeneity their usage by Semantic Web technologies and in combination with Linked Data is limited. For the Web context, Local-as-View (LAV) data integration paradigm is well suited [1], nevertheless the traditional techniques used to produce query answers, *query rewritings*, may be too expensive in the context of SPARQL queries and numerous sources [21].

Many of the Linked Data datasets provide SPARQL endpoints.[2] These endpoints allow users to explore datasets, and even use federated SPARQL query engines to answer queries using linked data across several datasets, i.e., *federated queries*. Unfortunately as the number and complexity of queries posed to the endpoints increases, their availability decreases [27]. An alternative to improve data availability is to give data consumers a more active role, and use their resources to replicate data, and consequently increase the data availability [18]. As Linked Data consumers are autonomous participants, the replication cannot closely follow the techniques used in distributed databases, and new strategies to select and localize sources are needed.

Data replication has been already used by data consumers that, not being able to rely on SPARQL endpoints with availability limitations, have set up their

[1] According to the 2014 report about the state of Linked Data Cloud available at http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

[2] 615 SPARQL endpoints are registered at http://datahub.io/ (July, 2015)

own SPARQL endpoints. These endpoints are mostly mirrors, and their use is restricted to local users. This replication strategy is limited by the user resources, and a user with a modest amount of resources may not set up endpoints for all datasets relevant for all the queries she may want to answer.

Linked Data Fragments (LDF) [27] has been proposed to exploit clients resources to relieve servers resources and improve availability, and Col-graph [15] has been proposed to use clients resources to replicate datasets fragments and improve data quality. We think Linked Data users may also benefit from replicating datasets fragments to execute federated queries; but in order to achieve this, they need to put their own replicated fragments at other users disposal.

Solving the availability limitations of Linked Data using data consumer resources, to replicate data and to make these data available to others, may lead to query processing performance concerns. For example, given popular datasets with many data consumers willing to set up endpoints with datasets subsets, how are these endpoints going to be used to execute a federated query? A very simple solution may be to declare all of these endpoints as part of the federation used by a federated query engine, but this simple solution may incur in very expensive execution times. For example, in [24], we give an example where two replicas of the relevant fragments leads to increase the execution time in two orders of magnitude for federated query engines FedX [26] and ANAPSID [2]. In order to properly exploit the benefits of replicated datasets fragments, we propose a source selection strategy that is aware of fragment replication able to enhance federated query processing engines. This idea of sharing replicated fragments among Linked Data consumers is new, and so there are no existing techniques that can perform a source selection aware of data replication. Even if techniques to detect data overlapping based on data summaries exists [25], applying them to scenarios with data replication will incur in expensive computations that are unnecessary in a replication scenario, will have accuracy limitations that may be overcome by tailored approaches, and will make greedy decisions that choose the best sources for a triple pattern without considering that the triple pattern would be executed in combination with other triple patterns.

In this thesis, our aim is to process SPARQL queries using views in two contexts. First, in the context of data integration, answer queries using views that describe sources as conjunctive SPARQL queries. And second, in the context of SPARQL federations of endpoints, answer queries using queries that describe the data fragments that have been replicated across SPARQL endpoints.

The research objective is to improve the query execution performance. This improvement may be measured in terms of amount of transferred data or answer throughput.

In particular, we want to answer the following research questions: *I*) When SPARQL is used as language for LAV global schema, can view loading outperform traditional LAV query rewriting techniques in query answering? *II*) How can the knowledge about the fragment replication be used to safely reduce the number of selected sources by federated query engines while keeping the same answer completeness? *III*) Does considering BGPs instead of just triple patterns

at source selection time produce better source selections that lead to less intermediate results?

The main challenges of this work are: *i*) Choosing the order in which views are loaded into the local graph that represents a partial instance of the global schema. *ii*) Keeping the size of transferred data low. *iii*) Keeping low the complexity of the containment computation among replicated fragments.

2 Proposed Approach

We propose two approaches to solve our problem: the SemLAV approach, and the FEDRA approach. SemLAV integrates heterogeneous data from Linked Data and Deep Web, following the Local-as-View (LAV) paradigm [1] to describe the setup, it uses a graph instance that integrates views instances, built during query execution, to answer users queries. FEDRA selects the sources to be contacted to execute each subquery in order to produce the query answer.

2.1 Formal Definition and Properties of the Approach

The SemLAV approach is based on the LAV paradigm, wrappers are used to transform data from Deep Web sources into the global schema. Given a SPARQL query Q on a set M of LAV views, SemLAV selects relevant views for Q and ranks them in order to maximize query results. Next, data collected from selected views are included into a partial instance of the global schema, where Q can be executed whenever new data is included. The SemLAV approach has the following properties: *1*) If the global schema instance includes the data from all the relevant views ranked by SemLAV, the query execution produces the same answer as traditional rewriting-based query processing approaches. *2*) The effectiveness of SemLAV is proportional to the number of covered rewritings. *3*) The view loading and query execution time linearly depends on the number of views loaded in the global schema instance. *4*) Answers may be produced incrementally if all the relevant views do not fit in memory.

The FEDRA approach is based on query containment and equivalence to prune relevant sources to retrieve data from, and a reduction to the set covering problem to produce as few subqueries as possible, sending to the endpoints subqueries composed by as many triple patterns as possible, and hopefully reducing the number of intermediate results to transfer from endpoints to the query engine. Given the set of the endpoints descriptions that constitute the federation, and a query, find a function D that for each query triple pattern returns the set of endpoints that need to be contacted in order to obtain a query answer as complete as possible while intermediate results are reduced. The FEDRA approach has the following properties: *1*) FEDRA source selection used in combination with query engines, produces at least as many answers as the query engines alone. *2*) FEDRA selects as few sources as possible for each triple pattern. *3*) FEDRA source selection aims to reduce the number of transferred tuples from endpoints to query engines.

2.2 Relationship between your approach and state-of-art approaches

Several approaches have been proposed for querying the Web of Data [2, 6, 12, 13, 19]. All these approaches assume that queries are expressed in terms of RDF vocabularies used to describe the data in the RDF sources; thus, their main challenge is to effectively select the sources, and efficiently execute the queries on the data retrieved from the selected sources. In contrast, SemLAV attempts to integrate data sources, and relies on a global schema to describe data sources and to provide a unified interface to the users. As a consequence, in addition to collecting and processing data transferred from the selected sources, SemLAV decides which of these sources need to be contacted first, to quickly answer the query. The Local-As-View paradigm for data integration allows to easily integrate new data sources; further, data sources that publish entities of several concepts in the global schema, can be naturally defined as LAV views. Query rewriters allow to answer queries against sources described using the LAV paradigm, some examples are: MCD-SAT [5], GQR [17], Bucket Algorithm [20], and MiniCon [11].

Recently, DAW, a source selection duplication aware approach for Linked Data has been recently proposed [25]. DAW relies on indexes with data summaries, that allow to measure the overlapping among sources and properly reduce the number of selected sources. Nevertheless, keeping up-to-date endpoints data summaries may be very expensive, and they cannot guarantee the correct overlapping detection. Moreover, DAW is a triple pattern wise approach, and data localities are not used to choose sources that may reduce the number of transferred tuples from endpoints to the query engines. FEDRA endpoints description are perfect data summaries, with no accuracy issues with a size independent of the data size, and that requires less updates that data summaries. FEDRA does take into account the data localities to produce as few subqueries as possible and reduce the size of intermediate results.

Linked Data Fragments (LDF) [27], a new publishing strategy that reduces the data publisher resources usage, and proposes a more active role for the data consumers. Like LDF, our approach seeks to use data consumers computational resources, and hopefully it may contribute to improve data availability of data publishers. Contrarily to LDF, FEDRA approach seeks to share the query execution among data consumers, and does not intend that each client communicates with the data publisher, and creates a cache of her queries datasets.

FedX [26] and ANAPSID [2] are state-of-art query engines for federations of SPARQL endpoints. Both, FedX and ANAPSID, send queries for triple patterns that should be evaluated in more than one endpoint, individually to the endpoint; and group into exclusive groups (FedX) or star-shaped groups (ANAPSID) the triple patterns that can be solely evaluated using one endpoint. We have extended FedX and ANAPSID query engines with FEDRA source selection strategy, and study FEDRA impact during query execution.

3 Implementation of the Proposed Approach

We use SPARQL conjunctive queries as views that describe sources contents. As in the Bucket Algorithm [20], relevant sources are selected per query subgoal, i.e., triple pattern. And the selected sources are ranked according to the number of query subgoals that are covered by their views. Loading first the views that cover more subgoals aims to produce answers as soon as possible. We expect to observe that using LAV paradigm to answer SPARQL queries allows to integrate heterogeneous data sources, and that the SemLAV approach produce more answers sooner than traditional query rewriting-based approaches. We will use variable substitution [10] to detect containment among fragments, and containment to detect equivalence of fragments [11]. Endpoints that have replicated equivalent fragments are considered as equivalent sources for that fragment. Fragment containment will be used to select the set of non-redundant relevant fragments for each triple pattern. Then, a second selection will be done at BGP level to reduce the number of subqueries to be sent to the endpoints, and in consequence, the number of transferred tuples. To further prune the selected sources, a set covering heuristic [16] will be used to determine the minimal set of endpoints that can be used to execute the subqueries. The approach implementation may be done stand-alone, having as input a query without SERVICE clauses, it produces a query with SERVICE clauses that delegate subqueries execution to remote endpoints; or existing federated query engines may be extended with FEDRA source selection strategy, to be used before the engines optimizations. We expect to observe that query engines enhanced with FEDRA incur in less intermediate results, and produce answers sooner.

The SemLAV approach has been formalized [21], and empirical tests have been performed [21, 9]. Some limitations of SemLAV implementation are that it lacks of strategies to overcome the memory limitations, and that views are loaded sequentially because the data store used did not allow parallel loading of data. The FEDRA source selection strategy for federations with replicated fragments has been proposed in [24], and some empirical tests have been performed [24]. Currently FedX and ANAPSID query engines have been extended to include the FEDRA source selection strategy. A limitation of the FedX extension is that it may increase the number of intermediate results in some cases, when one endpoint is selected for triple patterns that do not share variables.

4 Empirical Evaluation Methodology

The approaches will be used to execute queries in different setups, and they will be compared with alternative approaches.

For the SemLAV approach, the Berlin Benchmark [7] dataset generator will be used to generate a ten million triples dataset, and queries and views based on the ones proposed in [8] will be used. To stress the scalability of SemLAV and query rewriting, 476 views are used. SemLAV performance is compared to

the performance of three query rewriters: GQR [17], MCD-SAT [5], and MiniCon [11]. Query execution performance will be measured in terms of answer size, query execution time, amount of memory used, and answer throughput.

Our research hypothesis is that the SemLAV approach will produce a higher answer throughput than traditional query rewriting-based approaches, for LAV data integration in the context of Linked Data and the Deep Web. Currently, the SemLAV evaluation has been performed, and results are available at [21].

For the FEDRA approach, we will study two federated query engines: FedX and ANAPSID. As baseline we will consider the federated query engines source selection strategy. We will consider real and synthetic datasets of varying sizes from 72 thousand to 10 millions triples. Query generators will be used to produce more than 10,000 queries for each dataset. And we will use each dataset to setup a federation of ten endpoints, and simulate the execution of federated queries. The federations will contain endpoints with heterogeneous replicated fragments, and a sample of 100 random queries will be taken for each federation. For each query, the number of selected sources per triple pattern, source selection time, execution time, answer completeness, and number of transferred tuples will be measured. The obtained results will be statistically analyzed using the Wilcoxon signed rank test for paired non-uniform data.

In particular, we have the following research hypotheses: *1*) The FEDRA approach will select significantly less sources than federated query engines like FedX and ANAPSID. *2*) The FEDRA approach enhances federated query like FedX and ANAPSID, and reduces the number of intermediate results during query execution. *3*) The reductions that the FEDRA approach achieves in terms of number of selected sources and intermediate results, do not reduce the number of obtained answers.

Experiments with six federations and extended version of the query engines have already been performed, and they support our research hypotheses [24].

5 Lessons Learned, Open Issues and Future Directions

Query rewriting approaches are too stressed by the large number of triple patterns in SPARQL queries and the high number of sources in the Web, these characteristics prevents query rewriters to offer a practical query evaluation strategy for Linked Data. SemLAV uses query rewriters most basic information, *buckets*, to select relevant views, and ranks sources in a way that when k views have been loaded, they cover the maximal number of rewriting that can be covered with k views. The SemLAV approach implementation can be improved by loading views in parallel, and considering memory limits.

In federations of SPARQL endpoints with replicated fragments, federated query engines need to be enhanced with a source selection approach like FEDRA that allows to select sources for each triple pattern in a way that data transferred from endpoints to query engines is reduced. The strategy used by FEDRA gets excellent results when used with ANAPSID, but results are less good when used with FedX. ANAPSID does not send subqueries with Cartesian products

to the endpoints, but FedX may do it when triple patterns that do not share variables need to be executed in the same endpoint. Cartesian products may significantly increase the number of transferred tuples when compared to the evaluation of each triple pattern individually. These limitations may be overcome with a stand-alone implementation of the FEDRA approach that transforms a plain query into a query with query decomposition and source localization represented as SERVICE clauses that avoids Cartesian products. Unfortunately, such implementation has faced with query engines limited support for SERVICE clauses. In this direction, we are currently working in a version that do rewrite the query using SERVICE clauses, and use some heuristics to overcome current federated query engines limitations during execution.

Other extensions of FEDRA include the usage of cost functions, and the consideration of divergent fragments. Cost functions may be used to select the endpoints that satisfy the user criteria, e.g., in [23] public endpoint usage was reduced. Removing the assumption that all the endpoints fragments descriptions are up-to-date, endpoints may offer data with different levels of divergence with respect the current dataset, in the same direction as in [22].

Acknowledgments. This thesis is supervised by Pascal Molli and Hala Skaf-Molli. It has received contributions from Maria-Esther Vidal.

References

1. S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset, and P. Senellart. *Web Data Management*. Cambridge University Press, New York, NY, USA, 2011.
2. M. Acosta, M. Vidal, T. Lampo, J. Castillo, and E. Ruckhaus. ANAPSID: An Adaptive Query Processing Engine for SPARQL Endpoints. In Aroyo et al. [4], pages 18–34.
3. G. Antoniou, M. Grobelnik, E. P. B. Simperl, B. Parsia, D. Plexousakis, P. D. Leenheer, and J. Z. Pan, editors. *The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I*, volume 6643 of *Lecture Notes in Computer Science*. Springer, 2011.
4. L. Aroyo et al., editors. *ISWC 2011, Part I*, volume 7031 of *LNCS*. Springer, 2011.
5. Y. Arvelo, B. Bonet, and M.-E. Vidal. Compilation of query-rewriting problems into tractable fragments of propositional logic. In *AAAI*, pages 225–230. AAAI Press, 2006.
6. C. Basca and A. Bernstein. Avalanche: Putting the Spirit of the Web back into Semantic Web Querying. In A. Polleres and H. Chen, editors, *ISWC Posters&Demos*, volume 658 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.
7. C. Bizer and A. Schultz. The berlin sparql benchmark. *Int. J. Semantic Web Inf. Syst.*, 5(2):1–24, 2009.
8. R. Castillo-Espinola. *Indexing RDF data using materialized SPARQL queries*. PhD thesis, Humboldt-Universität zu Berlin, 2012.
9. P. Folz, G. Montoya, H. Skaf-Molli, P. Molli, and M.-E. Vidal. SemLAV: Querying Deep Web and Linked Open Data with SPARQL. In *ESWC: Extended Semantic Web Conference*, volume 476 of *The Semantic Web: ESWC 2014 Satellite Events*, pages 332 – 337, Anissaras/Hersonissou, Greece, May 2014.

10. C. Gutierrez, C. A. Hurtado, A. O. Mendelzon, and J. Pérez. Foundations of Semantic Web databases. *J. Comput. Syst. Sci.*, 77(3):520–541, 2011.
11. A. Y. Halevy. Answering queries using views: A survey. *VLDB J.*, 10(4):270–294, 2001.
12. A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data summaries for on-demand queries over linked data. In M. Rappa, P. Jones, J. Freire, and S. Chakrabarti, editors, *WWW*, pages 411–420. ACM, 2010.
13. O. Hartig. Zero-knowledge query planning for an iterator implementation of link traversal based query execution. In Antoniou et al. [3], pages 154–169.
14. B. He, M. Patel, Z. Zhang, and K. C.-C. Chang. Accessing the Deep Web. *Commun. ACM*, 50(5):94–101, 2007.
15. L. D. Ibáñez, H. Skaf-Molli, P. Molli, and O. Corby. Col-Graph: Towards Writable and Scalable Linked Open Data. In P. Mika et al., editors, *ISWC 2014, Part I*, volume 8796 of *LNCS*, pages 325–340. Springer, 2014.
16. D. S. Johnson. Approximation Algorithms for Combinatorial Problems. In A. V. Aho et al., editors, *ACM Symposium on Theory of Computing*, pages 38–49. ACM, 1973.
17. G. Konstantinidis and J. L. Ambite. Scalable query rewriting: a graph-based approach. In T. K. Sellis, R. J. Miller, A. Kementsietsidis, and Y. Velegrakis, editors, *SIGMOD Conference*, pages 97–108. ACM, 2011.
18. D. Kossmann. The state of the art in distributed query processing. *ACM Computer Survey*, 32(4):422–469, 2000.
19. G. Ladwig and T. Tran. Sihjoin: Querying remote and local linked data. In Antoniou et al. [3], pages 139–153.
20. A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In T. M. Vijayaraman, A. P. Buchmann, C. Mohan, and N. L. Sarda, editors, *VLDB*, pages 251–262. Morgan Kaufmann, 1996.
21. G. Montoya, L. D. Ibáñez, H. Skaf-Molli, P. Molli, and M.-E. Vidal. SemLAV: Local-As-View Mediation for SPARQL Queries. *Transactions on Large-Scale Data- and Knowledge-Centered Systems XIII*, pages 33–58, 2014.
22. G. Montoya, H. Skaf-Molli, P. Molli, and M.-E. Vidal. Fedra: Query Processing for SPARQL Federations with Divergence. Technical report, Université de Nantes, May 2014.
23. G. Montoya, H. Skaf-Molli, P. Molli, and M.-E. Vidal. Efficient Query Processing for SPARQL Federations with Replicated Fragments. Jan. 2015.
24. G. Montoya, H. Skaf-Molli, P. Molli, and M.-E. Vidal. Federated SPARQL Queries Processing with Replicated Fragments. In *The Semantic Web - ISWC 2015 - 14th International Semantic Web Conference*, Bethlehem, United States, Oct. 2015.
25. M. Saleem, A.-C. N. Ngomo, J. X. Parreira, H. F. Deus, and M. Hauswirth. DAW: Duplicate-AWare Federated Query Processing over the Web of Data. In H. Alani et al., editors, *ISWC 2013, Part I*, volume 8218 of *LNCS*, pages 574–590. Springer, 2013.
26. A. Schwarte, P. Haase, K. Hose, R. Schenkel, and M. Schmidt. FedX: Optimization Techniques for Federated Query Processing on Linked Data. In Aroyo et al. [4], pages 601–616.
27. R. Verborgh, M. V. Sande, P. Colpaert, S. Coppens, E. Mannens, and R. V. de Walle. Web-Scale Querying through Linked Data Fragments. In C. Bizer et al., editors, *WWW Workshop on LDOW 2014*, volume 1184 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.

Scaling Out Sound and Complete Reasoning for Conjunctive Queries on OWL Knowledge Bases

Sambhawa Priya

Department of Computer Science and Engineering, Lehigh University
19 Memorial Drive West, Bethlehem, PA 18015, USA
sps210@lehigh.edu

Abstract. One of the challenges the Semantic Web community is facing today is the issue of scalable reasoning that can generate responsive results to complicated queries over large-scale OWL knowledge bases. Current large-scale semantic web systems scale to billions of triples but many such systems perform no reasoning or rely on materialization. On the other hand, most state-of-the-art, sound and complete DL reasoners are main memory-based and fail when given ontologies that include enormous data graphs in addition to expressive axioms. Thus, until now, reasoning has been restricted to either limited expressivity or limited size of the data. The focus of this thesis is to develop a scalable framework to perform sound and complete reasoning on large and expressive data graphs for answering conjunctive queries over a cluster of commodity machines. In order to achieve our goal, we outline our approach to address the following challenges: partitioning large and expressive datasets across the cluster for distributed reasoning, and allocating reasoning and query-execution tasks involved in processing conjunctive queries to nodes of the cluster. We include evaluation results for our preliminary framework.

Keywords: distributed reasoning, partitioning, conjunctive queries, scalable framework

1 Problem Statement

Objective of the Research: This thesis focuses on how to scale out sound and complete reasoning for answering conjunctive queries over real-world ontologies with increasingly large data graphs over commodity clusters. Our goal is to design a framework that can scale to clusters of commodity machines for answering complex queries over large-scale knowledge bases characterized by small but expressive TBox and very large ABox. In order to reach this objective, we plan to address the following research questions: How to partition expressive datasets across the cluster such that traditional DL reasoners can answer subproblems independently at each partition? How to recombine these results to get sound and complete answers? How to co-locate partitions to reduce communication overheads? How to allocate reasoning and query-execution tasks involved in processing conjunctive queries to different nodes of the cluster? Note, we restrict

our application to handle queries about instances (i.e. ABox queries) and not the classes and properties (i.e. TBox queries) in the ontology.

Motivation and Relevance: A framework that allows reasoning for answering queries over large linked datasets can be a useful tool for finding insights from large graphs of data coming from diverse sources such as medicine and health-care, finance, social networks, government and the Internet of Things. Our proposed system can allow users to quickly answer questions of interest without creating application specific code or performing domain specific graph analyses.

Challenges and Opportunities: The main challenges we want to be able to address are as follows:

a) Partitioning large, highly networked and expressive datasets across the cluster for distributed reasoning to answer conjunctive queries: The partitioning approach [1] we used in our preliminary implementation [2] works only for $SHIF$ DL. This approach might be limited by power law and, in case of highly networked data, might generate very large partitions. The challenge is to extend the expressivity of the approach and generalize it to handle very expressive, highly networked and complex datasets.

b) Assigning the data-partitions and query-execution tasks to the nodes in the cluster such that the reasoning and query-execution tasks involved in processing the conjunctive queries are efficiently performed: Most existing works [3, 6] on scalable and distributed processing of conjunctive queries do not take reasoning into account or rely on materialization. On the other hand, existing works on parallel implementation of backward-chaining reasoning [7] do not address the challenge of computing conjunctive queries that involve DL reasoning. We want to address a combination of both of these issues in a scalable manner.

2 Proposed Approach

In order to address the challenges listed above, our proposed approach is as follows: Adopting a divide-and-conquer approach, we plan to explore partitioning of large and expressive datasets so that reasoning can be performed on subsets of data in a distributed environment where CPUs and memory of many commodity machines are harnessed. We plan to devise strategies to select partitions relevant to a query which, not only takes into account the property or the constant terms appearing in the query, but also incorporates the reasoning involved in answering that query. We plan to explore techniques to allocate the data partitions to nodes such that the load is evenly balanced across the cluster and the cost of communicating intermediate query results between the nodes is minimized. Time taken to perform expressive reasoning over semantic web datasets can be vulnerable to the order in which the query clauses are evaluated. In order to address this, we propose to implement and evaluate techniques for reasoning-aware query planning and develop strategies for allocating subqueries to compute nodes to perform distributed reasoning for answering conjunctive queries efficiently.

Related Work: A few authors [4, 5] have proposed approach for partitioning large OWL TBox based on the structure of class hierarchy and dependencies between the TBox elements. However, our focus is on such knowledge bases where the TBox is small enough to be replicated on each compute node of the cluster but the ABox is very large and requires partitioning across the cluster for scalable reasoning. Most of the previous work on parallel and distributed reasoning on semantic web datasets has been limited to forward chaining for less expressive logic such as RDFS, while our focus is on performing backward-chaining reasoning on rich description logics, beginning with $SHIF$ and pushing towards achieving a distributed system for $SHROIQ$. Oren et al. [8] combine parallel hardware with distributed algorithms to implement a system called MARVIN for scalable RDFS reasoning. Weaver et al. [9] derive an 'embarrassingly parallel' algorithm for materializing complete RDFS closure using C/MPI. WebPIE [10] has been used to compute the transitive closure of up to 100 billion triples, using the OWL Horst fragment (which is less expressive than $SHIF$). AllegroGraph reports that a prerelease has been tested on up to 1 trillion triples [12], but the reasoning is limited to little more than RDFS (subsumption, domain, range) plus inverses, sameAs, and transitivity. One of the few known systems to perform backward chaining (in combination with materialization) is QueryPIE [7], a parallel engine for OWL Horst reasoning that has scaled to 1 billion triples using an 8 machine cluster. However, they show query evaluation for only single pattern queries, not conjunctive queries. Triple stores [13, 14] build specialized indices and apply many join optimization techniques to improve processing of conjunctive queries expressed in SPARQL. Gurajada et. al. [6] developed a distributed triple store with novel join-ahead pruning technique for the distributed processing of SPARQL queries. However, none of these triple stores perform any reasoning. Mutharaju et. al. [11] present an approach for distributed reasoning for OWL 2 EL ontologies where the main reasoning task is classification where as our focus is on distributing DL reasoning for conjunctive queries where finding all the answers to a conjunctive query is the primary reasoning task.

3 Implementation of the Proposed Approach

Partitioning of large and expressive datasets: In our preliminary work [2], we utilize a partitioning technique for OWL Lite datasets proposed by Guo and Heflin [1]. Since this technique is restricted to $SHIF$ DL, we plan to explore how this partitioning technique can be extended to $SHOIN$ and $SHROIQ$. One idea is to use theory approximation [15] to create a SHIF approximation of a more expressive set of axioms. In particular, if this approximation is a lower-bound on the models of the original theory, then any logical consequences of the original theory will also be logical consequences of the approximation. Introducing such approximations may lead to unsoundness resulting in partitions that will include triples that do not need to be grouped together, resulting in larger than needed partitions. However, since sound and complete reasoners are used to perform reasoning over each partition, the inferred knowledge will still

be sound and complete. We will evaluate using datasets of different sizes and expressivity to determine the limitations of the approximation approach. The approach described in [1] may be limited by power law which is characteristic of real-world, large and networked data. We plan to analyze the partitionability of such datasets, taken from Linked Data and synthetic data sources to investigate the limitations of the technique and adapt it to handle such arbitrary data graphs. In [1], after the ontology axioms have been analyzed, assertional triples that have common subjects or objects and interrelated predicates are placed in a common partition. This implies that it is possible to scale-out partitioning by creating a MapReduce version of the algorithm where the key is generated from the triple's subject and/or object and predicate.

Distributing data partitions across the cluster: The data-partitions can be distributed across the cluster during the query execution time or prior to all queries. For query-time distribution, we can select partitions relevant to a given query and distribute them among the compute nodes that results in even load balancing and minimized cost of transferring intermediate join-results. We can sequentially load partitions on compute nodes using multiple threads to even out the disk-access cost. Query-time distribution can be inefficient when dealing with multiple concurrent queries. However, in our first implementation, we will implement a framework to handle only one query at a time, and later adapt it to support multiple concurrent queries. We can also distribute data partitions across the cluster prior to all queries. The fundamental question here is how to determine the best data allocation for a mix of unknown queries. We want to explore whether we can can use heuristics to create a reasonable allocation for a query mix that fits certain basic assumptions. For example, we know that partitions containing the same predicate are more likely to be relevant to the same query triple pattern and, hence, should be spread across nodes in order to maximize their utilization. When processing a query, we can identify if the load is unbalanced and depending upon the query-processing task allocation on the nodes, we can determine if shuffling a subset of partitions between certain nodes can achieve a better load balance.

Reasoning-aware query planning for distributed query execution: The order of query clause evaluation is critical to the query response time. Processing selective query triple patterns and joins early can reduce the volume of intermediate results and can reduce the query processing time. Typical query optimization algorithms use statistics about the data such as the number of triples matching a given predicate, the number of distinct subjects/objects for each predicate, the distribution of these values using histograms, and statistics on the joined triple patterns [16]. However traditional SPARQL query optimization does not consider that reasoning can produce results with significantly different cardinalities than those estimated from the raw data. We plan to compute reasoning statistics as part of data partitioning process. In the long term, heuristics from previous queries can be cached and used for query planning when statistics fall short. For executing these query plans on the cluster, we need to allocate the query plan components to different nodes such that the reasoning and join-execution tasks

can be efficiently performed over the cluster. We plan to implement a master-slave architecture where the master will create a reasoning-aware logical query plan and a corresponding physical query plan to map the data and query tasks to different slaves; and slaves nodes will process their respective physical query plans concurrently, interleaving sound and complete reasoning for triple patterns with distributed join processing using a message-passing protocol. We plan to study the impact of our distributed query-answering framework on query performance using different query patterns involving complex reasoning on large scale real-world and synthetic datasets.

Current Implementation: We developed a preliminary framework [2] for parallel reasoning on partitioned dataset where we first partition the knowledge base using the strategy developed by Guo and Heflin [1] and then the execute reasoning tasks on data-partitions in parallel on independent machines. We implemented a master-slave architecture that distributes an input query (expressed in SPARQL query language) to the slave processes on different machines. All slaves run in parallel, each performing sound and complete reasoning using a tableau-based reasoner (Pellet) to execute each subgoal of the given conjunctive query on its assigned set of partitions. As a final step, the master joins the results computed by the slaves. We use an off-the-shelf database to store the results of query subgoals and to perform the final join on the results of the conjuncts. However, we have identified a few drawbacks in our preliminary framework: each compute node performs reasoning on every partition assigned to it, irrespective of the partition's relevance to the subgoal; and the relational database becomes a bottleneck while inserting and joining large intermediate results. As described in the previous section, we plan to implement an improved architecture, which, for a given query, generates reasoning-aware query plan and distributes relevant data partitions and query processing tasks across the cluster. We plan to implement our own distributed join-processing framework that will utilize a message passing protocol.

4 Empirical Evaluation Methodology

General Strategy: We plan to evaluate both a) the partitioning strategies for large and expressive datasets with different connectivity patterns, and b) the performance of our distributed reasoning framework. We will vary the size of the data, the size of the query (i.e., the number of query triple patterns), the number of compute nodes, the partitioning approach, and query optimization strategies for distributed query processing.

Hypotheses:

1. Applying theory approximation to more expressive set of axioms, as discussed in Section 3, will result in coarser ABox partitions but will preserve independence of resulting partitions.
2. Loosely connected datasets with weak axioms will produce a large number of smaller partitions. Vice versa for highly connected datasets with rich axioms.

3. It is possible to scale-out the partitioning technique by exploiting the MapReduce version of the algorithm where the key of each triple is generated from it's predicate, subject and/or object.
4. There exist data distribution strategies that can improve average query performance on an unknown mixes of queries while making minimal assumptions about those query.
5. Selectivity statistics about query triple patterns and their joins can be computed during the partitioning process without materialization of all triples and these statistics can be utilized to construct more efficient query plans.
6. Our distributed reasoning framework will perform better with rich queries involving complex reasoning and more number of join triple-patterns than the queries involving weak axioms and fewer triple patterns.

In table 1, we list the metrics and the datsets that we plan to use in our evaluation. Since there is a dearth of synthetic and real-world datasets that have large-scale data graph with very expressive DL axioms and realistic queries that can be used for testing our framework, we would like to explore the creation of synthetic datasets with tunable expressivity and queries with wide range of properties with respect to number of conjuncts, selectivity and diameter of the query graph. We would also like to augment some real world datasets with handcrafted DL axioms to test our framework.

Table 1. Metrics and Datasets

Metrics
With respect to partitioning: partitioning time, the total number of partitions, the size of resulting partitions (minimum, median, mean, maximum), and the standard deviation of partition size. *With respect to distributed reasoning*: query response time, utilization of compute nodes, amount of communication.
Datasets
Synthetic Datasets: Lehigh University Benchmark (LUBM)[17], University Benchmark (UOBM) [18] *Real-world datasets and ontologies*: DBPedia, Yago, Barton, DBLP, LinkedMDB, and Linked Life Data, Billion Triple Challenge datasets.

Preliminary Evaluation Results and Lessons Learned: In our preliminary framework [2], we conducted the experiments on LUBM data (LUBM-50, LUBM-100 and LUBM-200) to evaluate performance of the partitioner and found that the partitioning system scales well. It was possible to create very small partitions (for LUBM-200 with 27.6M triples, the largest partition had fewer than 6800 triples). Our experiments on the preliminary parallel reasoning framework using up to 32 nodes demonstrates significant parallelism, with 32 nodes being 3.5x faster than 8 nodes (see figure 1). The speedups fall short of embarrassing parallelism, mostly due to the setup cost (time spent distributing the query to

the slaves) and performing join to get the final answers. More details on this evaluation can be found in [2].

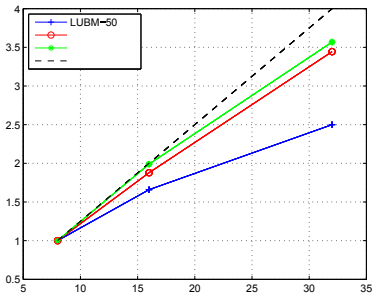

Fig. 1. Speed up for LUBM-50, LUBM-100 and LUBM-200 on 8, 16 and 32 processes.

5 Open Issues and Future Directions

Most existing scalable systems for processing conjunctive queries do not take reasoning into account. On the other hand, existing works on scalable reasoning are limited to forward chaining or reasoning on less expressive logic and do not handle conjunctive queries involving DL reasoning. In this thesis, we address both the issues by proposing a scalable and distributed framework for performing sound and complete reasoning for answering conjunctive queries over increasingly large data graphs involving expressive DL axioms. We plan to address the following core open issues in the future: a) partitioning large, highly networked and expressive datasets across the cluster for distributed reasoning, b) determining the best data allocation strategy for a mix of unknown queries such that load is evenly balanced and communication cost is minimized across the cluster, and c) assigning reasoning-aware query-plan components to the nodes for efficient processing of the reasoning and query-execution tasks involved in processing the conjunctive queries.

Stage of doctoral work: Middle.

Acknowledgement: I would like to thank Prof. Jeff Heflin (adviser) and Prof. Michael Spear (co-adviser) for their valuable comments on this paper.

References

1. Y. Guo and J. Heflin . A Scalable Approach for Partitioning OWL Knowledge Bases. In Proc. of the 2nd International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2006). Athens, Georgia. 2006.

2. S. Priya, Y. Guo, M. Spear, and J. Heflin. Partitioning OWL Knowledge Bases for Parallel Reasoning. In Eighth IEEE International Conference on Semantic Computing (ICSC 2014), Newport Beach, CA, 2014.
3. J. Huang, Daniel J. Abadi, and Kun Ren. Scalable SPARQL Querying of Large RDF Graphs. In Proceedings of Conference on VLDB. 4(11):1123-1134. 2011.
4. H. Stuckenschmidt and M. Klein. Structure-based partitioning of large concept hierarchies. In Proc. 3rd International Semantic Web Conference, Hiroshima, Japan, November 7-11, 2004.
5. A. Schlicht and H. Stuckenschmidt. A Flexible Partitioning Tool for Large Ontologies. In Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Sydney, Australia, December 9-12, 2008). IEEE Computer Society 482-488.
6. S. Gurajada, S. Seufert, I. Miliaraki and M. Theobald: TriAD: A Distributed Shared-Nothing RDF Engine based on Asynchronous Message Passing, Proceedings of the 2014 ACM International Conference on Management of Data (SIGMOD 2014), Snowbird, UT, USA, 2014.
7. J. Urbani, F. van Harmelen, S. Schlobach, H. E. Bal: QueryPIE: Backward Reasoning for OWL Horst over Very Large Knowledge Bases. International Semantic Web Conference (1) 2011: 730-745
8. E. Oren, S. Kotoulas, G. Anadiotis, R. Siebes, A. Teije, and F. van Harmelen. MARVIN: A platform for large scale analysis of Semantic Web data, In: Proceedings of the WebSci'09: Society On-Line, 18-20 March 2009, Athens, Greece.
9. J. Weaver and J. Hendler. Parallel materialization of the finite RDFS closure for hundreds of millions of triples, In Proceedings of the ISWC '09, 2009.
10. J. Urbani, S. Kotoulas, J. Maassen, F. Van Harmelen and Henri Bal. OWL reasoning with WebPIE: calculating the closure of 100 billion triples, Journal of Web Semantics, Vol 10, 2012.
11. R. Mutharaju, P. Hitzler, P. Mateti, and F. Lcu. Distributed and Scalable OWL EL Reasoning. In Proceedings of the 12th Extended Semantic Web Conference, Portoroz, Slovenia, To Appear, 2015.
12. Franz Inc. AllegroGraph RDFStore Benchmark Results, 2014. http://franz.com/agraph/allegrograph/agraph benchmarks.lhtml.
13. T. Neumann and G. Weikum. Scalable Join Processing on Very Large RDF Graphs. In Ugur C etintemel, Stanley B. Zdonik, Donald Kossmann, and Nesime Tatbul, editors, SIGMOD Conference, pages 627640. ACM, 2009.
14. A. Harth, J. Umbrich, A. Hogan, and S. Decker. YARS2: A Federated Repository for Querying Graph Structured Data from the Web. In ISWC/ASWC '07, Proceedings of the 6th international The semantic web and 2nd Asian conference on Asian semantic web conference, LNCS, volume 4825, pages 211224. Springer, 2007.
15. B. Selman and H. Kautz. Knowledge Compilation Using Horn Approximations. In Proceedings of Ninth National Conference on Artificial Intelligence (AAAI 1991), 1991, 904-909.
16. M. Stocker, A. Seaborne, A. Bernstein, C. Kiefer, and D. Reynolds. SPARQL basic graph pattern optimization using selectivity estimation. In Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008, pages 595604. ACM, 2008.
17. Y.Guo and Z.Pan and J.Heflin. LUBM: A Benchmark for OWL Knowledge Base Systems. Journal of Web Semantics, 3(2), 158-182, 2005.
18. L. Ma, Y. Yang, Z. Qiu, G. Xie, Y. Pan and S. Liu. Towards a complete OWL ontology benchmark. In Proceedings of the 3rd European conference on The Semantic Web (ESWC'06), pages 125-139, Springer, 2006.

Early Detection and Forecasting of Research Trends

Angelo Antonio Salatino

Knowledge Media Institute, The Open University, United Kingdom
angelo.salatino@open.ac.uk

Abstract. Identifying and forecasting research trends is of critical importance for a variety of stakeholders, including researchers, academic publishers, institutional funding bodies, companies operating in the innovation space and others. Currently, this task is performed either by domain experts, with the assistance of tools for exploring research data, or by automatic approaches. The constant increase of research data makes the second solution more appropriate, however automatic methods suffer from a number of limitations. For instance, they are unable to detect emerging but yet unlabelled research areas (e.g., Semantic Web before 2000). Furthermore, they usually quantify the popularity of a topic simply in terms of the number of related publications or authors for each year; hence they can provide good forecasts only on trends which have existed for at least 3-4 years. This doctoral work aims at solving these limitations by providing a novel approach for the early detection and forecasting of research trends that will take advantage of the rich variety of semantic relationships between research entities (e.g., authors, workshops, communities) and of social media data (e.g., tweets, blogs).

Keywords: Scholarly Data, Research Trends, Trend Detection, Trend Forecasting, Semantic Web Technologies.

1 Problem Statement

The research environment evolves rapidly: new potentially interesting research areas emerge regularly while others fade out, making it difficult to keep up with such dynamics. The ability to recognise important new trends in research and forecasting their future impact is however critical not just for obvious stakeholders, such as researchers, institutional funding bodies, academic publishers, and companies operating in the innovation space, but also for any organization whose survival and prosperity depends on its ability to remain at the forefront of innovation.

Currently, the task of understanding what the main emergent research areas are and estimating their potential is usually accomplished by experts with the help of a number of systems for making sense of research data. Systems such as Google Scholar, FacetedDBLP [1] and CiteSeerX [2] provide good interfaces which allow users to find scientific papers, but they do not directly support identification of research trends. Other tools such as Microsoft Academic Search, Rexplore [3], Arnetminer [4], and Saffron [5] provide a variety of visualizations that can be used for trend analysis, such as publication trends and co-authorship paths among researchers. However, the manual detection of research trends is an intensive and time-consuming task. Moreover, the constant increase in the number of research data published every year makes the approach based on human experts less and less feasible. It is thus important to

develop automatic and scalable methods to detect emerging research trends and estimate their future impact.

Currently, there are a number of approaches for detecting topic trends in a fully automatic way [6,7]. These are usually based on the statistical analysis of the impact of certain labels associated with a topic. However, these tools are unable to take full advantage of the variety of research data existing today and need to examine a significant number of years (e.g., 3-4) before they are able to identify and forecast topic trends [8,9]. In addition, they are only able to identify topics that have been explicitly labelled and recognized by researchers [10]. However, it can be argued that a number of topics start to exist in an embryonic way, often as a combination of other topics, before being officially named by researchers. For example, the Semantic Web emerged as a common area for researchers working on Artificial Intelligence, WWW and Knowledge-Based Systems, before being recognized and labelled in the 2001 paper by Tim Berners-Lee et al. [11].

The doctoral work presented here aims to solve the aforementioned limitations and produce a novel approach to detect and forecast research topics. This approach will be based on two main intuitions. First, I believe that by analysing the various dynamics of research it should be possible to detect a number of patterns that are correlated with the creation of new embryonic topics, not yet labelled. For example, the fact that a number of authors from previously unrelated research communities or topics are starting to collaborate together may suggest the emergence of a new interdisciplinary research area. Secondly, I theorize that taking into account the rich variety of semantic relationships between research entities (e.g., authors, workshops and communities) and analysing their diachronic evolution, it should become possible to forecast a topic impact in a much shorter timescale, e.g., 6-18 months. This holistic and semantic-based analysis of the research environment is today made possible by the abundance of both scholarly data and other sources of evidence about research, including social networks, blogs, and so on.

2 Relevancy

In many real-world contexts, being aware of research dynamics can bring significant benefits. **Researchers** need to be updated regularly on the evolution of research environments because they are interested in new trends related to their topics and potentially interesting new research areas. For **academic publishers** or **editors** knowing in advance new emerging topics is crucial for offering the most up to date and interesting contents. For example, an editor can gain a competitive advantage by being the first one to recognize the importance of a new trend and publish a special issue or a journal about it. **Institutional funding bodies** and **companies** need also to be aware of research developments and promising research trends. Thus, an automatic approach to detect novel topics and estimate their potential will bring significant advantages to a variety of stakeholders. Indeed support for this PhD project comes from Springer-Verlag, which is a global publishing company.

3 Related work

Several tools and approaches for the exploration of scholarly data already exist. From the perspective of topic trend detection, we can classify these systems as either semi-automatic or fully automatic. In particular, some systems for exploring the publication space provide implicit support for semi-automatic trend detection, such as Google Scholar, FacetedDBLP [1] and CiteSeerX [2]. Other systems offer instead an explicit support for semi-automatic trend detection, like Arnetminer [4], Microsoft Academic Search (MAS), Saffron [5] and Rexplore [3]. However, while all these systems are able to identify and visualize historical research trends, they do not provide any support for the detection of future ones.

In the context of providing a fully-automatic way to detecting topic trends, many approaches assess the impact of a topic by simply using the number of publications or patents directly associated with it. For example, Wu et al [8] integrate bibliometric analysis, patent analysis and text-mining analysis in order to detect research trends. Some models also take in consideration the citation graph. For example, Bolelli et al. [6] propose an author-topic model to identify topic evolution and then they use citations to evaluate the weight for the main terms in documents. He et al. [7] combine Latent Dirichlet Allocation and citation networks for detecting topics and understand their evolution. However, these approaches are able to detect trends only after the associated research areas are already established and they do not provide any support to the early detection of research trends.

State of the art methods for forecasting trends in research take usually into consideration the number of publications and authors associated with a topic [12], or the probability distribution of a topic over time [13]. They then analyse these time series either by means of statistical techniques [10] or machine learning methods [14], yielding a prediction for the following years. However, these methods do not take advantage of the knowledge that can be extracted by analysing the dynamics of multiple research entities (e.g., communities, venues), and they ignore the growing mass of research data that today can be acquired from social networks.

Another important aspect that needs to be taken into account is how to represent a topic. In literature, several ways to define a topic model can be found. The first is characterised by the use of keywords as proxies for research topics. Systems like MAS and Saffron [5] use this kind of model. This approach has several drawbacks because it does not take in consideration the relationships among research topics [15] and keywords tend to be noisy. The second kind of approach is the probabilistic topic model. Latent Dirichlet Allocation [16], which treats a document as a mixture of topics and a topic as a distribution over words, is the most popular of these methods. However, this model assumes that the topics used to generate a document are uncorrelated, which may be a risky assumption for research topics [17]. Other approaches for probabilistic topic model try to deal with this problem introducing a separability condition [18]. A third solution is using an explicit semantic topic model [9,17,3], which exploits a semantic network of research areas linked by semantic relations. The advantage of this solution is that it goes beyond the use of noisy, uncorrelated keywords and exploits instead an ontology of research areas.

4 Research Questions

Considering the gaps identified in the previous section, the main research question of the PhD will be: *"How is it possible to detect the early emergence of new research topics and forecast their future impact?"*.
This question entails two different challenges. The first one is how to detect very early research topics that may not even be labelled. The second one is how to forecast their impact with good accuracy. A specific set of sub-questions has been articulated in order to describe the process through which the doctoral work plans to answer the questions above.

Q1 – Finding the data. Understanding which data to integrate and exploit for the process is the first step. In particular, it is important to investigate the value of non-scholarly data (e.g., tweets, blogs, micro-posts, slides) in supporting trend detection and forecasting. As far as semantic technologies are concerned: how can research elements be gathered and connected by means of semantic relations?

Q2 – Detection of new emerging research topic. How is it possible to extract patterns in the evolution of research areas in order to predict the emergences of new ones? How can historical patterns be used to support the detection of future trends? Is it possible to develop a general approach able to consider the peculiarities of different fields (e.g., Computer Science, Business, Medicine and so on)? How emerging and unnamed research areas can be labelled? How social media can contribute in the detection of research trends?

Q3 – Forecasting of research trends. Can the impact of a research topic be measured just in terms of number of citations and publications? As soon as it has been defined, how the impact of research areas can be forecasted? What kind of forecasting approach should be adopted for research areas that do not yet exist? Which contribution can be given from the social media?

5 Hypotheses

From a philosophical point of view, academic disciplines are specific branches of knowledge which together form the unity of knowledge that has been produced by the scientific endeavour. When two or more disciplines start to cooperate they share their theories, concepts, methods and tools. The results of this cooperation may lead either to the creation of a new interdisciplinary research area or simply to a contribution in knowledge from one area to another. The basic hypothesis is that the creation of a topic is thus anticipated by a number of dynamics involving a variety of research entities, such as other topics, research communities, authors, venues and so on. Therefore, recognizing these dynamics might enable a very early detection of emerging topics.

Scholarly data can be used to analyse a huge amount of research elements such as papers, authors, affiliations, venues, topic and communities [19]. All these research elements are inherently interconnected by relations that can be defined as either explicit or implicit. Figure 1 shows, as an example, the six basic explicit connections between the research elements according to our model.

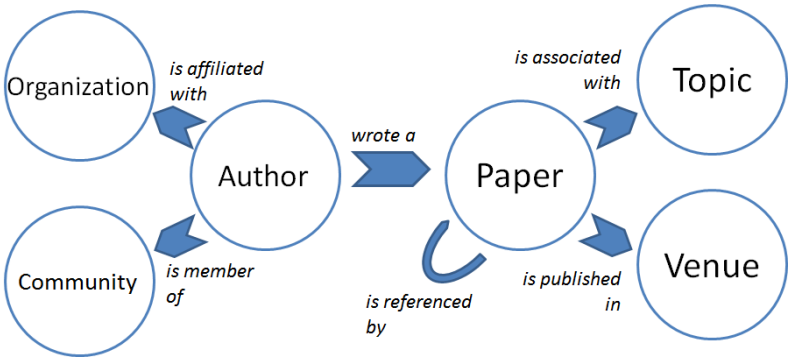

Fig. 1. Model representing the scholarly meta-data and their relationships

These explicit connections can be used to derive a number of second order connections, e.g. a topic is also associated with publication venues through relevant papers published in venues. These relationships can be analysed diachronically to derive the dynamics that led to the emergence of a topic and to estimate how they affect its future impact. For example, if two communities start to share research interests or authors, this may lead to the fact that a common new topic is developing. In a nutshell, the fundamental hypothesis at the basis of this PhD is that by exploiting the large variety of scholarly data which are now available, as well as modelling their semantic relationships, it will be possible to perform detection and forecasting of research trends even in a relative small interval of time. In addition, since many researchers are actively involved on social networks, I believe that analysing data from social media can also provide an effective support for the detection of research dynamics.

6 Approach

The approach is structured according to the proposed research questions. Basically, it is organised in four main steps.

Data integration. In this first phase I plan to integrate a variety of heterogeneous data sources, including both scholarly metadata and less traditional sources of knowledge, such as tweets, blogs post, slides and so on. The output will be a comprehensive knowledge base containing both the research entities from Figure 1 and entities from social media (authors' profiles, number of followers, analytics, etc.). I will identify topics and communities by extending state of the art techniques. In particular, I plan to treat topics semantically, by describing their relationships using the topic networks produced by the Klink algorithm [17]. I am also planning to use the approach for detecting topic-based research communities described in [19], since it explicitly links communities and topics.

The rich network of semantic relationship between the research elements will be described by an ontology and it will be populated by semi-automatic statistical methods. To build it, I plan to extend the topic network created by Klink with the research entities discussed in section 5 and their relationships. The analysis of these relationships and how they change in time will support the next steps of the approach.

Exploration of the Research Dynamics. In this step, the dynamics involving research elements correlated with the emergence of new topics will be investigated. To do so, I plan to verify empirically a number of hypotheses about these dynamics. In particular, I will analyse a number of topics which appear in the 2000-2010 interval and verify if their emergence is correlated with a number of dynamics, such as the raise of co-publications of related research areas, the increase of collaborations between authors of related areas, shifts of interests or migration phenomena in related communities, transfer of topics between related venues, and so on.
The output of this analysis will be a collection of patterns of knowledge flows associated with the creation of a new research area.

Early topic detection. This step aims to exploit the previously defined patterns for early research trend detection. To this end, I will build a number of distinct graphs, in which nodes represent a kind of research entity (e.g., topics) and the links are one of the elements of the dynamics, which were found in the previous phase – e.g., the increase in the number of collaborations between authors from two distinct topics. Highly connected sub-graphs, representing the area in which multiple entities exhibit the identified dynamics could thus suggest that a new discipline is emerging. In order to produce more robust evidence, I will use the semantic network of research entities to confirm that the emergence of a new topic is supported by a number of different 'traces' and research entities. For example, if a set of topics suggests that a correlated research area is emerging, the dynamics of the set of communities and venues related to these topics will also be checked. The intuition is that, while the evidence coming from a single dynamics or a single kind of entity could be biased or noisy, their combination should yield a more accurate result. The result will be a number of sets of linked entities, each one anticipating the emergence of a new topic. Different kinds of combination of entities and metrics will be tested, aiming to find the best approach to derive sets that are strongly correlated with the creation of new topics. At this stage, another challenge will be the definition of a method for labelling future research topics.

Trend forecasting. Initially, I will investigate different techniques to estimate the impact of a topic, taking in consideration both basic metrics, such as the number of publications and citations, and more complex indexes. As mentioned before, in contrast with current approaches, [8,9], I aim to develop a method which will be able to work also on relatively short time series (6-18 months). In order to do so, I will take advantage of a wide variety of features associated with a topic, representing both the performances of related entities (e.g., the track record of significant authors) and the previously discussed dynamics. Hence, I will conduct a comprehensive analysis of the correlations between these features and the topic impact in the following years. For example, I will analyse how the performance of related authors, communities, workshops, hashtags, scientific opinion leaders, and so on, influence on the previously defined impact metrics. It is hypothesised that such abundance and diversity of the features will compensate for the small interval of time in which early topics will be analysed. Moreover data from the social web and other real-time information, such as the number of views and downloads on the publisher sites and open access reposito-

ries, will offer a more granular timeline for the analysis of the topics, measured in weeks, rather than in years.

A set of different machine learning methods, such as Artificial Neural Networks, Support Vector Machines and Deep Belief Networks, will exploit the extracted features in order to forecast the performance of a topic.

7 Evaluation plan

I plan to conduct an iterative evaluation during the different phases of my work using both quantitative and qualitative approaches.

From a quantitative point of view, I will evaluate both the ability of the system to identify novel topics and its accuracy to assess their impact in the following years. The discussed approaches will be compared with current methods and the difference between their performances will be measured via statistical tests. I will evaluate the detection of emerging trends in terms of recall, precision and F-measure using cross-validation on historical data. Similarly, I will assess the agreement between the estimated and the real impact of a research area.

In the qualitative evaluation, the achieved results will be compared with experts' opinions in order to measure its reliability. I will prepare a number of surveys for domain experts with questions both about the past - such as the main topics recently emerged in their area of expertise - and about the future - such as the research areas which seem on the verge of being created and an estimation of their likely impact.

8 Conclusions

This paper presents the goal of my doctoral work, which is currently at an early stage (month 6). As discussed, I intend to produce a new approach for detecting and forecasting research trends, which is based on a semantic characterization of research entities, on the statistical analysis of research dynamics and on the integration of scholarly and social media data.

Currently I am investigating a number of knowledge sources for selecting the ones more apt to support my approach. At the same time I am using an initial dataset to test the hypotheses about research dynamics discussed in section 6. The next step will be the creation of an approach for extracting highly connected sub-graphs of entities exhibiting dynamics associated with the emergence of new topics.

Acknowledgements

I would like to acknowledge my advisors, Enrico Motta and Francesco Osborne, as well as my sponsors, Springer-Verlag GMBH for supporting my research.

References

1. Diederich J, Balke W-T, Thaden U Demonstrating the semantic growbag: automatically creating topic facets for faceteddblp. In: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries, 2007. ACM, pp 505-505

2. Li H, Councill I, Lee W-C, Giles CL CiteSeerx: an architecture and web service design for an academic document search engine. In: Proceedings of the 15th international conference on World Wide Web, 2006. ACM, pp 883-884
3. Osborne F, Motta E Rexplore: Unveiling the dynamics of scholarly data. In: Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on, 2014. IEEE, pp 415-416
4. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD, 2008. ACM, pp 990-998
5. Monaghan F, Bordea G, Samp K, Buitelaar P Exploring Your Research: Sprinkling some Saffron on Semantic Web Dog Food. In: SW Challenge - ISWC, 2010. Citeseer, pp 420-435
6. Bolelli L, Ertekin Ş, Giles CL (2009) Topic and trend detection in text collections using latent dirichlet allocation. In: Advances in Information Retrieval. Springer, pp 776-780
7. He Q, Chen B, Pei J, Qiu B, Mitra P, Giles L Detecting topic evolution in scientific literature: how can citations help? In: Proceedings of the 18th CIKM, 2009. ACM, pp 957-966
8. Wu F-S, Hsu C-C, Lee P-C, Su H-N (2011) A systematic approach for integrated trend analysis—The case of etching. Technological Forecasting and Social Change 78 (3):386-407
9. Decker SL, Aleman-Meza B, Cameron D, Arpinar IB (2007) Detection of bursty and emerging trends towards identification of researchers at the early stage of trends. (Doctoral dissertation, University of Georgia).
10. Tseng Y-H, Lin Y-I, Lee Y-Y, Hung W-C, Lee C-H (2009) A comparison of methods for detecting hot topics. Scientometrics 81 (1):73-90
11. Berners-Lee T, Hendler J, Lassila O (2001) The semantic web. Scientific american 284 (5):28-37
12. Budi I, Aji RF, Widodo A (2013) Prediction of Research Topics on Science & Technology (S&T) using Ensemble Forecasting. International Journal of Software Engineering and Its Applications 7 (5):253-268
13. Zhou D, Ji X, Zha H, Giles CL Topic evolution and social interactions: how authors effect research. In: Proceedings of the 15th CIKM '06, 2006. ACM, pp 248-257
14. Jun S, Uhm D (2010) Technology forecasting using frequency time series model: Bio-technology patent analysis. Journal of Modern Mathematics and Statistics 4 (3):101-104
15. Osborne F, Motta E, Mulholland P (2013) Exploring scholarly data with rexplore. In: The Semantic Web–ISWC 2013. Springer, pp 460-477
16. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. the Journal of machine Learning research 3:993-1022
17. Osborne F, Motta E (2012) Mining semantic relations between research areas. The Semantic Web-ISWC 2012, pp 410-426
18. Arora S, Ge R, Halpern Y, Mimno D, Moitra A, Sontag D, Wu Y, Zhu M (2012) A practical algorithm for topic modeling with provable guarantees.
19. Osborne F, Scavo G, Motta E (2014) Identifying diachronic topic-based research communities by clustering shared research trajectories. In: The Semantic Web: Trends and Challenges. Springer, pp 114-129

Entity Disambiguation for Wild Big Data Using Multi-Level Clustering

Jennifer Sleeman

Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore. MD 21250 USA
jsleem1@cs.umbc.edu

Abstract. When RDF instances represent the same entity they are said to corefer. For example, two nodes from different RDF graphs[1] both refer to same individual, musical artist James Brown. Disambiguating entities is essential for knowledge base population and other tasks that result in integration or linking of data. Often however, entity instance data originates from different sources and can be represented using different schemas or ontologies. In the age of Big Data, data can have other characteristics such originating from sources which are schema-less or without ontological structure. Our work involves researching new ways to process this type of data in order to perform entity disambiguation. Our approach uses multi-level clustering and includes fine-grained entity type recognition, contextualization of entities, online processing of which can be supported by a parallel architecture.

Introduction

Often when performing knowledge base population, entities that exist in the knowledge base need to be matched to entities from newly acquired data. After matching entities, the knowledge base can be further enriched with new information. This matching of entities is typically called entity disambiguation (ED) or coreference resolution when performed without a knowledge base [15]. Early work related to record linkage [6] was foundational to this concept of entity similarity. Though there is a significant amount of research in this area including methods which are supervised and unsupervised, these approaches tend to make assumptions that do not hold for big data.

Existing research tends to assume a static batch of data, ignoring the streaming, temporal aspects. It assumes that the schemas or ontologies are available and complete. Often issues such as heterogeneity and volume are not considered. However, big data applications tend to include unalignable data from multiple sources and often have schemas or ontologies that are absent or insufficient. We define these characteristics in terms of 'Wild Big Data' (WBD) [21] and describe how these characteristics challenge the disambiguation process. Our work specifically addresses these characteristics with an approach that could be used to perform ED for WBD.

[1] http://dbpedia.org/resource/James_Brown and http://musicbrainz.org/artist/20ff3303-4fe2-4a47-a1b6-291e26aa3438#

Objective

The objective of this research is to perform ED given the data is large in volume, potentially schema-less, multi-sourced and temporal by nature. We want to answer questions such as, how do we perform ED in a big data setting, can we efficiently distribute the task of ED without a loss in precision, how do we account for data that is changing over time, how do we process semantic graphs given they may not have an associated schema or ontology, and finally how do we process this data given it originates from different sources with potentially unalignable vocabularies. These questions are important to answer because they are real problems in big data applications [25].

Motivation

Big data is a growing area of research and offers many challenges for ED [2, 11]. The main motivation of this work is the need for ED that supports data with big data characteristics. This includes data originating from different sources which contain different types of entities at different levels of granularity, data that may not have a schema or ontology, and data that changes over time. This sort of data at big data volumes complicates the ED process.

Companies, organizations and government entities are sharing more data and acquiring more data from other sources to gain new insight and knowledge [8]. Often the combination of sources, such as social media, news and other types of sources can provide more insight into topics than a single source.

As is evident by efforts related to Linked Open Data (LOD) [17], interoperability among different data sources is of growing importance and essential for sharing data. As more data is made available for sharing, the need for aligning schemas/ontologies is increasing.

Knowledge bases typically contain entities, facts about the entities and links between entities. As new data is made available over time, these knowledge bases require ways to manage new information such as adding entities, links and new attributes pertaining to the entities. There is a need to also alter existing information such that information that becomes invalid over time is adjusted. For example, a link may become invalid or an attribute may prove to be incorrectly assigned to an entity.

Challenges and Opportunities

By exploring how to perform ED for big data, we will offer a strong contribution to this area as previous research has only focused on various parts of this problem.

Regarding the LOD [17], interoperability is a real challenge, particularly because vocabularies are not always alignable. For example, *address* in one vocabulary could mean *street address* alone and in another it could include *city*, *state* and *zip code*. We explored this problem in more depth in our previous work [18]. LOD attempts to provide a way for data providers to link their data into the cloud. However data may not always be made available as LOD, and in order for an application to perform ED, this alignment becomes essential.

With unstructured text, one can use natural language processing to acquire various facts related to entities found in the text. With RDF data, an ontology can often be used to develop an understanding of the data. However, when data is semi-structured such as RDF or JSON and no such ontology or schema is present, disambiguating entities becomes problematic. Making sense of these large data extractions becomes a real issue.

Knowledge bases naturally change over time, however it is a challenge to enrich the knowledge base over time while at the same time reducing errors in previously asserted facts. Algorithms used to perform ED are typically developed for static data. Incremental updates and changes are harder to incorporate. However, this is precisely what is needed as often big data applications are producing data on a periodic basis. If one is developing a knowledge base where facts are changing over time, the ED algorithm must accommodate these changes in a way that does not require the algorithm to reprocess all potential matches given new information.

Volume requires that the algorithm can be distributed in such a way that work could be performed in parallel. Again ED algorithms do not typically assume data in terms of the volume that is present with big data applications. However, since ED algorithms have typically $O(n^2)$ complexity, distributing the algorithm would be necessary for such large volumes of data.

Recent research which has addressed big data ED has primarily been in the natural language processing domain. For example, a number of researchers [4, 13, 16] have explored using MapReduce for pairwise document similarity. However, they are primarily focused on the volume characteristic. Work by Araujo et al. [1] tackled the problem of working with heterogeneous data but they worked with sources where the vocabularies were alignable. Work by Hogan et al. [9] addresses this problem of performing ED for large, heterogeneous data. However, they assume they have access to the ontologies used and they assume they can make use of owl:sameAs semantics (which isn't always present). The hard problem of trying to understand data absent knowledge of how it is structured has not been thoroughly addressed in previous research.

Proposed Approach

We are developing a multi-level clustering approach that includes one level of topic modeling and a second level of clustering using our own custom algorithm. This approach makes big data ED more tractable. Our research makes three major research contributions that work together to achieve an effective approach for performing online ED.

Research Contribution: Fine-grained Entity Type Recognition: If we consider identifying traits of an entity, at the highest level of identification, entities are defined by types, for example "Person", "Football Player", "Baseball Stadium", etc. With TAC [2] there are just three types used (PER, ORG, GEP) used, with DBpedia there are fewer than 1000 types, and tens of thousands of types in Yago. Given a WBD data set, data can contain a mix of entities,

[2] http://www.nist.gov/tac

can be composed of many different types, such as a person, a sports player, a team member, and can be defined by types that are defined at different levels of granularity. For example, "Person" is at a much higher level than "Football Player". Often type information is not available, to get around this problem, we have proposed a solution [21] based on topic modeling that enables us to define types when type information is not present.

Research Contribution: Multi-dimensional Clustering: We are developing a clustering algorithm that performs ED based on multiple dimensions. This algorithm would be applied to the coarse clusters generated from the fine-grained entity type recognition. The complexity of clustering algorithms can range from $O\left(n^2\right)$ to $O\left(n^3\right)$, so a key aspect of this work is that it supports parallelism.

Research Contribution: Incremental Online modeling to support Temporal Change: Our work includes knowledge base (KB) population. When entities are assessed as similar, the information in the KB is merged with the information contained in the newly recognized matched entity instance. However, similarity is usually associated with some level of probability. As more data is acquired over time, previous assertions may prove to have a lower probability than previously asserted.

Relationship with State of the art

As it relates to coreference resolution the following work [12, 1, 24] would be considered state of the art and is comparable to our work. In the NLP domain, a number of researchers have focused on scalable entity coreference using MapReduce [4, 13, 16].

As it relates to type identification, work by Ma et al. [10] presents a similar problem, whereby type information is missing. This work builds clusters that represent entity types based on both schema and non-schema features. Paulheim et al. [14] also address the problem of identifying type information when it is non-existent and they also use their approach to validate existing type definitions. They take advantage of existing links between instances and assume that instances of the same types should have similar relations. They acquire this understanding by examining the statistical distribution for each link.

As it relates to candidate selection, the following work [23, 15] would be considered state of the art and comparable to our work.

Implementation of Proposed Approach

We will implement our approach as a software system by which it could be used to perform ED for wild big data. We will demonstrate the effectiveness of this approach by launching it in a parallel environment processing wild big data. We will use benchmarks to convey the overall performance of the system as it compares to other systems that are not necessarily addressing the wild big data aspects. We anticipate ED scores that have slightly lower precision but we expect to see better computing performance as we scale the number of entities in our system to big data sizes, since our approach is developed to be amenable to a Hadoop-like architecture.

Current Implementation

For the first level of clustering we use Latent Dirichlet Allocation (LDA) [3] topic modeling, to form coarse clusters of entities based on their fine-grained entity types. We use LDA to map unknown entities to known entity types to predict the unknown entity types. We shared preliminary results of this effort in our previous work [21]. Table 1 shows our latest results that include experiments using DBpedia data where we show accuracy given we found all types and accuracy given we missed on 1 type but found the others. Figure 1 also shows another experiment where we measured precision at N where given N predictions we found all of the types for a particular entity. This approach offers two benefits, it results in overlapping clusters based on entity types improving recall and it does not require knowledge of the schema or ontology of the data. The only requirement is that there is a knowledge base of entity types that can be used as a source for associating entity types to unknown entities.

We have performed research related to ED of people in our early work [19] where we experimented with combining rules and supervised classification, however when ED is performed on entities of different types in combination with different data sources, the ED process is more difficult. Often recognizing the types of entities and then performing ED among specific types can reduce this problem, however, when the data sets are large to the scale of big data problems, even recognizing these types reduces the problem to intractable sized subproblems. For this reason, we are building a custom clustering algorithm for the second level of clustering. This work is still in-process.

We have performed preliminary work [20] with hierarchical clustering and did not find this to be a viable solution. Our current work clusters based on a number of features such as distance measures, co-occurrences, graph-based properties, and statistical distributions. Distinctive to our work, we also incorporate context which we derive from our topic model. Entity context provides additional information about an entity that is not necessarily acquired from the associated predicates for that entity. We are also currently performing preliminary experiments related to contextualizing entities.

Current Limitations

Since our approach is a two-level approach, errors from the first level of clustering could propagate to the second level. We look to overcome this problem by generating a model that both levels of clustering would use, however a resolution to this problem is still under investigation.

This approach is currently limited to graph-based data. There is a lot of unstructured text and it would be advantageous for our system to be able to convert unstructured text to graph-based structures. In addition, in order for our approach to work with data that is truly "wild", we require access to a knowledge base that is rich with fine-grained entity types. The richness of the knowledge base and its representation of the data to be processed directly influence how well our approach will perform. For example, if our knowledge base has very little information related to car accidents and we are processing entities from

a data source related to car accidents, we will under-perform when recognizing the fine-grained entity types which consequently will negatively impact our ED algorithm.

Empirical Evaluation Methodology

Since there are multiple parts to our approach, we intend to evaluate the various parts in addition to how well the parts work together to perform ED.

Hypotheses

1. By using a multi-level clustering approach we can perform ED for wild big data and achieve F-measure rates that are close to those of other ED algorithms that are not processing wild big data.
2. Fine-grained entity type recognition as a first level of clustering is a competitive approach to performing candidate selection.
3. Our approach will be scalable such that it is comparable with other methods that perform ED in parallel.
4. By performing ED online, we can reduce the number of errors in our KB.

General Strategy

Our general approach for evaluation is to evaluate our first level of clustering, the fine-grained entity type recognition work in isolation of ED. We will then perform experiments related to contextualizing entities, performing ED both from a scalability and accuracy perspective, and finally online KB improvements.

Benchmarks We will use data sets that we are able to easily establish ground truth for, such as DBpedia and Freebase. However, we will also use Big Data datasets and we may use unstructured data sets that are processed by an OpenIE [5] system resulting in triple-based information.

Our goal with the fine-grained entity type recognition work is to be able to identify all entity types that are assigned to gold standard entities. We also will try to identify incorrectly used and missing entity types. We will perform experiments which benchmark our first level of clustering with other candidate selection methods and will be benchmarked against an existing type identification approach [14].

With our second level of clustering we hope to demonstrate that contextualization of entities improves performance. We also plan to compare our ED with others from an accuracy standpoint and from a complexity standpoint. We will benchmark how well we scale in a parallel environment compared to other parallel ED approaches.

One feasible approach for evaluating the ED method is to use data from the LOD and remove links then compare our results with the unlinked data to the data that is linked [7]. We will also explore the Instance Matching Benchmark

[3] for evaluation and benchmarking. Another benchmark that is more recent is SPIMBench [4] which provides test cases for entity matching and evaluation metrics, and supports testing scalability. Finally we will show how a KB with online temporal changes can reduce errors over time. We will prove this by taking an offline KB and comparing it to our online version.

Metrics For our evaluation we will use the standard F-measure metric. For evaluating our clusters, we will likely use standard clustering metrics such as measuring purity.

$Precision = \frac{TruePositive}{TruePositive+FalsePositive}$
$Recall = \frac{TruePositive}{TruePositive+FalseNegative}$
$F - measure = 2 * \frac{Precision*Recall}{Precision+Recall}$

Current State of Evaluation

Our early work [22] shows our evaluation of identifying fine-grained entity types using an entropy-based approach. We now use a topic modeling approach and have performed preliminary evaluation of this work [21]. We also include in Table 1 our latest evaluation. This evaluation is based on DBpedia 6000 randomly selected entities and 176 types used to build the model. We used 350 separately randomly selected entities that are of type *Creative Works*, type *Place*, and type *Organization*, as these had the highest representation among the training set. We measured how often we were able to recognize all types associated with each entity as defined by DBpedia. We are also in the process of a comprehensive evaluation for this work. We are currently developing our custom clustering algorithm and will plan to evaluate this work soon. We performed preliminary experiments with an online KB where we reduced the errors by 70% by updating the KB over time.

Table 1: Fine-Grained Entity Type Accuracy

Test	Avg Num Types	Accuracy (0 Types Missed)	Accuracy (1 Type Missed)
CreativeWork	6	.76	.91
Place	7	.60	.67
Organization	9	.74	.77

Lessons Learned, Open Issues, and Future Directions

One of our challenges is finding the data we need to properly evaluate our approach. Since we are proposing a system that works with Big Data scale datasets, our evaluations will be harder to achieve.

A second challenge is comparing and benchmarking our work against others. Since our approach addresses problems that may overlap with other research

[3] http://islab.dico.unimi.it/iimb/
[4] http://www.ics.forth.gr/isl/spimbench/index.html

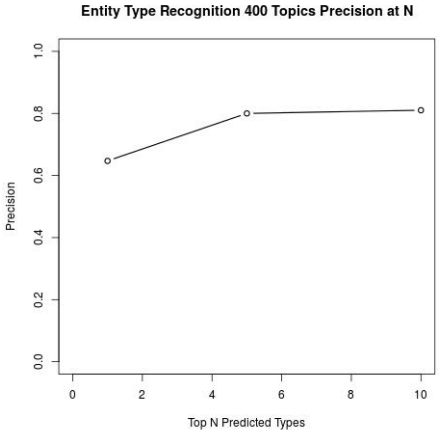

Fig. 1: Fine-Grained Entity Type Precision at N

but isn't exactly the same, we will need to benchmark parts of our system with other research.

From our previous experiments when evaluating mappings of entity types from one data source to another we learned that since there will not always be a direct mapping, we will need to have supporting heuristics which makes the evaluation process harder to achieve. For example mapping between Freebase and DBpedia is not always possible, often because types defined in one knowledge base just do not exist in the other.

Acknowledgments

The author would like to thank her advisor, Dr. Tim Finin, and Dr. Anupam Joshi.

References

1. Araujo, S., Tran, D., DeVries, A., Hidders, J., Schwabe, D.: Serimi: Class-based disambiguation for effective instance matching over heterogeneous web data. In: WebDB. pp. 25–30 (2012)
2. Beheshti, S.M.R., Venugopal, S., Ryu, S.H., Benatallah, B., Wang, W.: Big data and cross-document coreference resolution: Current state and future opportunities. arXiv preprint arXiv:1311.3987 (2013)
3. Blei, D.M.: Probabilistic topic models. Communications of the ACM 55(4), 77–84 (2012)
4. Elsayed, T., Lin, J., Oard, D.W.: Pairwise document similarity in large collections with mapreduce. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers. pp. 265–268. Association for Computational Linguistics (2008)
5. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM 51(12), 68–74 (2008)

6. Fellegi, I.P., Sunter, A.B.: A theory for record linkage. Journal of the American Statistical Association 64(328), 1183–1210 (1969)
7. Ferrara, A., Montanelli, S., Noessner, J., Stuckenschmidt, H.: Benchmarking matching applications on the semantic web. In: The Semanic Web: Research and Applications, pp. 108–122. Springer (2011)
8. Franks, B.: Taming the big data tidal wave: Finding Opportunities in Huge data streams with advanced Analytics, vol. 56. John Wiley & Sons (2012)
9. Hogan, A., Zimmermann, A., Umbrich, J., Polleres, A., Decker, S.: Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. Web Semantics: Science, Services and Agents on the World Wide Web 10, 76–110 (2012)
10. Ma, Y., Tran, T., Bicer, V.: Typifier: Inferring the type semantics of structured data. In: Data Engineering (ICDE), 2013 IEEE 29th Inter. Conf. on. pp. 206–217. IEEE (2013)
11. McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D., Barton, D.: Big data. The management revolution. Harvard Bus Rev 90(10), 61–67 (2012)
12. Nikolov, A., Uren, V., Motta, E., Roeck, A.: Overcoming schema heterogeneity between linked semantic repositories to improve coreference resolution. In: Proc. 4th Asian Conf. on the Semantic Web. vol. 5926, pp. 332–346 (December 2009)
13. Pantel, P., Crestan, E., Borkovsky, A., Popescu, A.M., Vyas, V.: Web-scale distributional similarity and entity set expansion. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. pp. 938–947. Association for Computational Linguistics (2009)
14. Paulheim, H., Bizer, C.: Type inference on noisy rdf data. In: International Semantic Web Conference (2013)
15. Rao, D., McNamee, P., Dredze, M.: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, Multilingual Information Extraction and Summarization, pp. 93–115. Springer (2013)
16. Sarmento, L., Kehlenbeck, A., Oliveira, E., Ungar, L.: An approach to web-scale named-entity disambiguation. In: Machine Learning and Data Mining in Pattern Recognition, pp. 689–703. Springer (2009)
17. Schmachtenberg, M., Bizer, C., Paulheim, H.: State of the LOD cloud. http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/ (2014)
18. Sleeman, J., Alonso, R., Li, H., Pope, A., Badia, A.: Opaque attribute alignment. In: Proc. 3rd Int. Workshop on Data Engineering Meets the Semantic Web (2012)
19. Sleeman, J., Finin, T.: Computing foaf co-reference relations with rules and machine learning. In: The Third Int. Workshop on Social Data on the Web. ISWC (November 2010)
20. Sleeman, J., Finin, T.: Cluster-based instance consolidation for subsequent matching. Knowledge Extraction and Consolidation from Social Media p. 13 (2012)
21. Sleeman, J., Finin, T.: Taming wild big data. In: Symposium on Natural Language Access to Big Data (2014)
22. Sleeman, J., Finin, T., Joshi, A.: Entity type recognition for heterogeneous semantic graphs. In: AI Magazine. vol. 36, pp. 75–86. AAAI Press (March 2105)
23. Song, D., Heflin, J.: Automatically generating data linkages using a domain-independent candidate selection approach. In: Int. Semantic Web Conf. (2011)
24. Song, D., Heflin, J.: Domain-independent entity coreference for linking ontology instances. Journal of Data and Information Quality (JDIQ) 4(2), 7 (2013)
25. Suchanek, F., Weikum, G.: Knowledge harvesting in the big-data era. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 933–938. ACM (2013)

Profiling the Linked (Open) Data

Blerina Spahiu

Università degli Studi di Milano-Bicocca
spahiu@disco.unimib.it

Abstract. The number of datasets published as Linked (Open) Data is constantly increasing with roughly 1000 datasets as of April 2014. Despite this number of published datasets, their usage is still not exploited as they lack comprehensive and up to date metadeta. The metadata hold significant information not only to understand the data at hand but they also provide useful information to the cleansing and integration phase. Data profiling techniques can help generating metadata and statistics that describe the content of the datasets. However the existing research techniques do no cover a wide range of statistics and many challenges due to the heterogeneity nature of Linked Open Data are still to overcome. This paper presents the doctoral research which tackles the problems related to Linked Open Data Profiling. We present the proposed approach and also report the initial results.

Keywords: Linked Open Data, Profiling, Data Quality, Topical Classification

1 Problem Statement

With 12 datasets in 2007, the Linked Open Data cloud has grown to more than 1000 datasets as of April 2014 [17], a number that is constantly increasing. The datasets to be published need to adopt a series of rules in a way that it would be simple for them to be searched and queried [3]. The datasets should be published adapting W3C standarts in RDF[1] format and made available for SPARQL[2] endpoint queries. Adapting these rules allow different data sources to be connected by typed links which are useful to extract new knowledge as linked datasets do not have the same information. Even though the Linked Open Data is considered a gold mine, its usage is still not exploited as understanding a large and unfamiliar RDF dataset is still a key challenge. As a result of a lack of comprehensive descriptive information the consumption of these dataset is still low. Data profiling techniques support data consumption and data integration with statistics and useful metadata about the content of the datasets. While traditional profiling techniques solve many issues these techniques can not be applied to heterogeneous data such as Linked Open Data. Data profiling techniques in the context of Linked Open Data are very important for different tasks:

[1] http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
[2] http://www.w3.org/TR/rdf-sparql-query/

Complex schema discovery. Schema complexity leads to difficulties to understand and access databases. Schema summaries provide users a concise overview of the entire schema despite its complexity.

Ontology / schema integration. Ontologies published on the Web, even for datasets in similar domains can have differences. Data profiling techniques can help understanding the overlap between ontologies and help in the process of ontology creation, maintenance and integration.

Big knowledge bases and provide a landscape view. Data profiling techniques can help identifying some core knowledge patterns (KP) which reveal a piece of knowledge in a domain of interest.

Inspect large datasets to find quality issues. Data profiling tools allow the inspection of large datasets for detecting quality issues, by identifying the cases that do not follow business rules, outliers detection, residuals, etc.

Data integration. To perform a data integration process, one should consider schema mapping, the process of discovering relationships between schemas. Profiling techniques can reveal mappings between classes and properties, helping the integration process.

Entity summarization. Finding features that best represent the topic/s of a given dataset can help not only the topical classification of the dataset but also understanding the semantic of the information found in the data.

Data visualization for summarization. Profiling techniques can support data visualization tools to visualize large multidimensional datasets by displaying only a small and concise summary of the most relevant and important features, enhancing the comprehension of the user by allowing him to dig into the data by zooming in or out the provided summary.

In this proposal we will focus on the profiling techniques to summarize the content of a dataset and reveal data quality problems. Moreover we will propose profiling techniques combined with data mining algorithms to find useful and relevant features to summarize the content of datasets published as Linked Open Data and also techniques that reveal quality issues in the data. The dataset summarization can be used not only to detect if the dataset is useful or not, but also to provide useful information to the cleansing and integration phase.

2 Related Works

Statistics and summaries can help to describe and understand large RDF data. Most of the existing profiling tools, support traditional databases which are homogeneous and have a well-defined schema. These techniques can not be applied to Linked Open Data due to their heterogeneity and the lack of a well-defined schema. As it will be discussed most of the existing techniques to profile Linked Open Data are limited in few statistics and summaries covering only one task.

Roomba [1] is a framework to automatically validate and generate descriptive dataset profiles. The extracted metadata are grouped into four categories (general, access, ownership or provenance) depending on the information they hold. After metadata extraction some validation and enrichments steps are performed.

Metadata validation process identifies missing information and automatically corrects them when it is possible. As an outcome of the validation process, a report is produced which can be automatically sent to the dataset maintainer.

The ExpLOD [8] tool is used to summaries a dataset based on a mechanism that combines text labels and bisimulation contractions. It considers four RDF usages that describe interactions between data and metadata, such as class and predicate instantiation, class and predicate usage on which it creates RDF graphs. It also uses the `owl:sameAs` links to calculate statistics about the interlinking between datasets. The ExpLOD summaries are extracted using SPARQL queries or algorithms such as partition refinement.

RDFStats [9] generates statistics for datasets behind SPARQL endpoint and RDF documents. It is built on Jena Semantic Framework and can be executed as a stand-alone process, important to optimize SPARQL queries. These statistics include the number of anonymous subjects and different types of histograms; URIHistogram for URI subject and histograms for each property and the associated range(s). It uses also methods to fetch the total number of instances for a given class, or a set of classes and methods to obtain the UIRs of instances.

LODStats [2] is a profiling tool which can be used to obtain 32 different statistical criteria for datasets from Data Hub. These statistics describe the dataset and its schema and include statistics about number of triples, triples with blank nodes, labeled subjects, number of `owl:sameAs` links, class and property usage, class hierarchy depth, cardinalities etc. These statistics are then represented using Vocabulary of Interlinked Datasets (VoID)[3] and Data Cube Vocabulary[4].

ProLOD [5] is a web based tool which analyzes the object values of RDF triples and generates statistics upon them such as data type and patterns distribution. In ProLOD the type detection is performed using regular expression rules and normalized patterns are used to visualize huge numbers of different patterns. ProLOD also generates statistics on literal values and external links. ProLOD++[5] which is an extension of ProLOD is also a browser based tool which implements several algorithms with the aim to compute different profiling, mining or cleansing tasks. In the profiling task are included processes to find frequencies and distribution of distinct subjects, predicates and objects, range of the predicates etc. ProLOD++ can also identify predicates combinations that contain only unique values as key candidates to distinctly identify entities. The implementation of mining tasks cover processes such as synonym and inverse predicate discovering, association rules on subjects, predicates and objects, etc. It also performs some cleansing tasks such as auto completions of new facts for a given dataset, ontology alignment in identifying predicates which are synonym or identifying cases where the pattern usage is over specified or underspecified.

Profiling as the activity of providing insights through the data, is not only about providing statistics about value distribution, null values etc, but also is referred to the process of finding and extracting information patterns in the data.

[3] http://www.w3.org/TR/void/
[4] http://www.w3.org/TR/vocab-data-cube/
[5] https://www.hpi.uni-potsdam.de/naumann/sites/prolod++/app.html

In the area of schema summarization Knowledge Patterns (KP) can be defined as a template to organise meaningful knowledge [6]. The approach in [15] identifies an abstraction named dataset knowledge architecture that highlights how a dataset is organized and which are the core knowledge patterns (KP) we can retrieve from that dataset. These KPs summarise the key features of one or more datasets, revealing a piece of knowledge in a certain domain of interest.

Encyclopedic Knowledge Patterns (EKP) [12] are some knowledge patterns introduced to extract core knowledge for entities of a certain type from Wikipedia page links. EKPs are extracted from the most representative classes describing a concept and containing abstraction of properties. The use of EKPs that supports exploratory search is showen in Aemoo[6] to enrich query results with relevant knowledge coming from different data sources in the Web [13].

In order to understand complex datasets, [4] introduces Statistical Knowledge Pattern (SKP) to summarize key information about an ontology class considering synonymity between two properties of a given class. An SKP is stored as an OWL ontology and contains information about axioms derived or not expressed in a reference ontology but can be promoted applying some statistical measures.

As shown, the actual profiling tools provide schema based statistics like the class/property usage, incoming/outgoing links etc, but none of the existing works is focused in providing summarization of the content of the dataset and also apply techniques to profile its quality. Author in [7] propose an approach to profile the Web of Data, but in difference from this, the proposed approach profiles Linked Data in terms of its quality and summarize datasets in terms of its topic.

3 Research Plan

The contribution of this PhD in the area of Linked Open Data Profiling covers (i) generating new statistics that are not covered by the state of the art techniques (ii) new algorithms to overcome the challenges to perform profiling in the LOD, and (iii) the development of a methodology on how to perform profiling tasks. In the following we will give an overview of the methodology which we want to follow in order to accomplish the contribution we want to make in the field.

New statistics for Linked Data Profiling

While much effort is done as described in the state of the art, the generated statistics are limited in some basic statistics such as the number of triples, number of classes/ properties that are used in a dataset, the datatypes or `sameAs` links used, etc. Datasets hold much more interesting information which might be hidden, but at the same time, this information could be useful for the consumer of the dataset. As data profiling is referred to the activity of providing useful descriptive information, new techniques on how to extract the hidden information should be developed. Our intent is to develop automatic approaches to generate new statistics and knowledge patterns to provide dataset summary and inspect its quality. Different data mining techniques, such as association rule mining, can be used to discover and extract patterns and dependencies in the dataset. These

[6] http://wit.istc.cnr.it/aemoo/

patterns might provide useful information especially to detect errors and inconsistencies in spatial data (*consistency* quality dimension). Implementation of different approaches for outlier detection, like distance/deviation/depth-based, evolutionary techniques, etc. could provide insight about abnormalities in the underlying data. Other techniques such as clustering, classification, aggregation, dimensionality reduction or spatial data summarization might help to provide concise and accurate dataset summarization and inspect quality dimensions mentioned in [16]. We intend to further investigate the topical classification of LInked Open Data. The datasets published as LOD cover a wide range of topics but they lack metadata that describe the topical category, so the users have difficulties deciding if the dataset is relevant for their interest or not. For each of the dataset published as LOD a label for the topical category was manually assigned [17]. The datasets have only one label for the topical category while often two or more topics are needed to describe a dataset. The actual topical classification of datasets in the LOD is limited to eight categories, while a more fine-grained topical classification might provide more useful information.

Overcoming Profiling Challenges

As another contribution in this research we want to tackel the profiling challenges described in [11]. Traditional profiling task can not be applied to Linked Data due to their heterogeneity. Heterogeneity can appear in different forms such as different formats or query languages called syntactic heterogenity. Linked Open Data can be represented in different formats, stored in different storage architectures also the data encoding schemes may vary. This is referred to as schematic heterogeneity. Datasets published as LOD might use different vocabularies, to describe synonymous terms. [11] referred semantic heterogeneity as the discovery of semantic overlap of the data. Traditional data profiling tools can not be used to profile Linked Open Data as they suppose data to be homogeneous stored in a single repository, while Linked Open Data are neither homogeneous nor stored in a single repository. Also as the number of the datasets published is increasing the need to adapt and optimise profiling techniques to support huge amount of data is also high. A good approach when dealing with large datasets, is to improve the profiling performance running the calculation of statistics and patterns extraction in parallel. We also plan to adapt some data mining techniques to deal with high dimensionality data, such as Linked Open Data.

Methodology to Profile Linked Open Data

As another contribution of this research we intend to develop a methodology on how to perform profiling tasks. This methodology would classify profiling tasks depending on the purpose and also provide guidelines to appropriate select the tasks needed by the user.

4 Preliminary Results

This PhD work is now at the second year. As a first step we measured the value of Linked Open Data, profiling the data published as Open Data from the Italian Public Administrations. In this work we profiled the adoption of Linked Open

Data best practices and local laws by the Italian Public Administration calculating a compliance index considering three quality dimensions for the published data; *completeness, accuraccy* and *timeliness* [18].

As mentioned in the Sec. 3, the main contribution of this research is to provide new techniques for dataset summarization and new statistics about the data. ABSTAT[7] is a framework which can be used to summarise linked datasets and at the same time to provide statistics about them. The summary consists of Abstract Knowledge Patterns (AKPs) of the form <*subjectType, predicate, objectType*> which represent the occurrence of triples <sub,pred,obj> in the data, such that subjectType is a minimal type of sub and objectType is a minimal type of obj. The ABSTAT summaries can help users comparing in which of two datasets a concept is described with richer and diverse properties, and also help detecting errors in the data such as missing or datatype diversity, etc [14]. ABSTAT can also be used to fix the domain and range information for properties. Either the domain or the range is unspecified for 585 properties in DBpedia Ontology and AKPs can help us in determining at least one domain and one range for the unspecified properties. For example, for the property `http://dbpedia.org/ontology/governmentType` in DBpedia we do not have information about the domain. With our approach we can derive 7 different AKPs meaning that we can derive 7 domains for this property.

We further investigated one of the challenges still present in the Linked Open Data datasets, topic classification. We built the first automatic approach to classify LOD datasets into the topical categories that are used by the LOD cloud diagram. For the classification we considered eight feature sets; vocabulary, classes and properties usage, local class/property names, text from `rdfs:lable`, top-level domain and in and out degree. In Table 1, are shown the results training three classifiers k-NN, Naive Bayes and Decision Tree on three balancing approaches, no sampling, down and up sampling and two normalization techniques considering the binary occurrence and the relative term occurence for each term or vocabulary. Our approach achieves an accuracy of 81,62% [10].

Table 1. Results of combined feature sets. Best three results in bold.

Classification Approach	Accuracy in %			
	ALL_{bin}	ALL_{rto}	NoLab_{bin}	NoLab_{rto}
k-NN (no sampling)	74.93	71.73	76.93	72.63
k-NN (down sampling)	52.76	46.85	65.14	52.05
k-NN (up sampling)	74.23	67.03	71.03	68.13
J48 (no sampling)	**80.02**	77.92	79.32	79.01
J48 (down sampling)	63.24	63.74	65.34	65.43
J48 (up sampling)	79.12	78.12	79.23	78.12
Naive Bayes (no sampling)	21.37	71.03	**80.32**	77.22
Naive Bayes (down sampling)	50.99	57.84	70.33	68.13
Naive Bayes (up sampling)	21.98	71.03	**81.62**	77.62

A deep literature study for the tools which are used to profile LOD has been taken. We analyzed existing tools in terms of the goal they are used for,

[7] http://abstat.disco.unimib.it/

techniques, input, output, approach, automatization information, license etc, with the aim to have a complete view of the existing approaches and techniques for profiling which helps us in determining new statistics or new techniques. This deepen study will also help us for the third contribution classifying profiling tasks and creating a general methodology for each task depending on the use case.

5 Lessons Learned, Open Issues and Future Work

The main contribution of this PhD work is to address the challenges mentioned in Sec. 3 to built a framework for profiling the Linked Open Data in order to give insights of the data, despite their heterogeneous nature. To evaluate the validity of the proposed approach or the results achieved is very difficult as in the filed of LOD profiling there is no Gold Standard, thus is very difficult to compare with others. For this issue, we want to further explore how these new statistics or summarization allow to improve the performance of the actual profiling techniques and tools, e.g. how profiling tasks can improve full-text search etc. To evaluate the validity of the proposed profiling techniques to summarise datasets, as pattern discovery is not trivial, humans will evaluate the validity of the summarization in terms of relatedness and informativeness. We intend to provide to users a list of statistics and ask them which in their opinion is more important to support profiling of Linked Open Data. The evaluation of the performance of profiling tasks is very difficult, which still remains an open issue on which I am currently working.

The ABSTAT framework provides some contributions in summarising Linked Open Data, and detecting quality issues. We are working to enrich this framework with other statistics and to apply it to unstructured data such as microdata.

Regarding the topical classification of LOD datasets, we will consider the problem for multi-label classification. As the datasets in the LOD cloud are unbalanced a two stage approach might help, while a classifiers chain which makes a prediction for one class after the other could address the multi-lable problem. Up till now in our experiments we have not exploited RDF links beyond datasets in and out degree, so link-based classification techniques could be applied to further investigate the content of a dataset.

Acknowledgements

This research has been supported in part by FP7/2013-2015 COMSODE (under contract number FP7-ICT-611358). I would like to thank my supervisor Assoc. Prof Andrea Maurino, my supervisor during my visiting period Prof. Dr Christian Bizer, Asst. Prof Matteo Palmonari, Dr. Anisa Rula for their priceless suggestions and also the anonymous reviewers for their helpful comments.

References

[1] A. Assaf, R. Troncy, and A. Senart. Roomba: An extensible framework to validate and build dataset profiles. In *The 2nd International Workshop on Dataset PRO-*

FIling and fEderated Search for Linked Data (PROFILES '15) co-located with ESWC 2015, Portorož, Slovenia, May 31 - June 1, 2015., pages 32–46, 2015.

[2] S. Auer, J. Demter, M. Martin, and J. Lehmann. Lodstats - an extensible framework for high-performance dataset analytics. In *18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012.*

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[4] E. Blomqvist, Z. Zhang, A. L. Gentile, I. Augenstein, and F. Ciravegna. Statistical knowledge patterns for characterising linked data. In *Proceedings of the 4th Workshop on Ontology and Semantic Web Patterns co-located with ISWC 2013, Sydney, Australia, October 21, 2013.*

[5] C. Böhm, F. Naumann, Z. Abedjan, D. Fenz, T. Grütze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with prolod. In *Workshops Proceedings of the 26th ICDE 2010, March 1-6, 2010, Long Beach, California, USA*.

[6] A. Gangemi and V. Presutti. Towards a pattern science for the semantic web. *Semantic Web*, 1(1-2):61–68, 2010.

[7] A. Jentzsch. Profiling the web of data. *Proceedings of the 8th Ph. D. retreat of the HPI research school on service-oriented systems engineering*, page 101, 2014.

[8] S. Khatchadourian and M. P. Consens. Explod: Summary-based exploration of interlinking and RDF usage in the linked open data cloud. In *ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010*, pages 272–287, 2010.

[9] A. Langegger and W. Wöß. Rdfstats - an extensible RDF statistics generator and library. In *Database and Expert Systems Applications, DEXA, International Workshops, Linz, Austria, August 31-September 4, 2009*, pages 79–83, 2009.

[10] R. Meusel, B. Spahiu, C. Bizer, and H. Paulheim. Towards automatic topical classification of lod datasets. In *Proceedings of the 24th International Conference on World Wide Web, LDOW Workshop, 2015, Florence,Italy, May 18-22, 2015.*

[11] F. Naumann. Data profiling revisited. *SIGMOD Record*, 42(4):40–49, 2013.

[12] A. G. Nuzzolese, A. Gangemi, V. Presutti, and P. Ciancarini. Encyclopedic knowledge patterns from wikipedia links. In *The Semantic Web - ISWC 2011 Bonn, Germany, October 23-27, 2011, Proceedings, Part I*, pages 520–536, 2011.

[13] A. G. Nuzzolese, V. Presutti, A. Gangemi, A. Musetti, and P. Ciancarini. Aemoo: exploring knowledge on the web. In *Web Science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2-4, 2013*, pages 272–275, 2013.

[14] M. Plamonari, A. Rula, R. Porrini, A. Maurino, B. Spahiu, and V. Ferme. Abstat: Linked data summaries with abstraction and statistics. In *European Semantic Web Conferenze 2015 (ESWC2015) Portoroz, Slovenia, 31th May - 4th June 2015.*

[15] V. Presutti, L. Aroyo, A. Adamou, B. A. C. Schopman, A. Gangemi, and G. Schreiber. Extracting core knowledge from linked data. In *Proceedings of the COLD 2011, Bonn, Germany, October 23, 2011*, 2011.

[16] A. Rula and A. Zaveri. Methodology for assessment of linked data quality. In *Proceedings of the 1st Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany, September 2nd, 2014.*, 2014.

[17] M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *The Semantic Web - ISWC 2014, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, pages 245–260, 2014.

[18] G. Viscusi, B. Spahiu, A. Maurino, and C. Batini. Compliance with open government data policies: An empirical assessment of italian local public administrations. *Information Polity*, 19(3-4):263–275, 2014.

Inferencing in the Large Characterizing Semantic Integration of Open Tabular Data

Asha Subramanian

International Institute of Information Techmology,
26/C, Electronics City, Hosur Road,, Electronic City, Bengaluru, Karnataka 560100, India
asha.subramanian@iiitb.org
www.iiitb.ac.in

Abstract. *Tables are a natural and ubiquitous way of representing related information. Actionable insight is usually gleaned from tabular datasets through data mining techniques assisted by domain experts. These techniques however, do not harness the semantics or the contextual reference underlying the datasets. Tabular datasets, especially the ones created as part of open data initiatives often contain information about entities fragmented across several datasets implicitly connected through some semantics thus giving them a contextual reference. Our work deals with harnessing this context (Thematic Framework) in which they can be reasoned further. This thesis aims at creating algorithmic support for a human to semantically integrate a collection of tabular data using ontologies from publicly available knowledge bases in Linked Open Data. The overall objectives of our work called "Inferencing in the Large" aims to go further than this, to enrich the mapped ontology with inferencing rules and generate enriched RDF (Schematic Framework), to enable the use of semantic reasoners.*

Keywords: Semantic Web · Linked Open Data · Context Abduction · Ontology · Graph Models

1 Introduction

Recent initiatives like Open Data by various governments, have resulted in a number of freely available tabular datasets containing actionable knowledge that could be relevant to several stakeholders. However, these published datasets are often created independently, with no overarching purpose or schematic structure. Indeed, there may be no overarching *thematic* structure – the datasets need not be about any one particular topic or theme. As a result, valuable knowledge remains fragmented across the datasets. While typical analytics efforts use data mining or machine learning algorithms to exploit data patterns and generate inferences, they fail to harness the implicit meaning of the data to make meaningful inferences in a semantic context.

There is a pressing need for *semantic integration* of such arbitrarily structured data.

Given a collection of tables, it is a daunting task even to determine what the set of tables are collectively about (that is, a topic or theme for the collection), let alone establish an overarching schematic framework.

We call this problem, "Inferencing in the Large" (in the wilderness), where in order to extract meaning from a collection of data, we first need to establish a framework within which meaning can be interpreted.

In a previous project called Sandesh (expanding to Semantic Data Mesh), we had proposed a knowledge representation framework based on Kripke Semantics called Many Worlds on a Frame (MWF) to integrate disparate datasets [1]. In this model, aggregated knowledge was represented in the form of several semantic "worlds" – each of which represented a schematic framework within which data was organized. Given a set of tabular data, and a hand-crafted set of seed worlds and their (type and location) relationships, the Sandesh toolkit reorganized the tabular data into data elements within the schematic frameworks of one or more worlds. MWF was meant to address the absence of a schematic and thematic framework in open datasets. However, the Sandesh framework still requires significant human effort in organizing the seed set of worlds and their schematic structures.

In this work, we aim to automate this process further, by using ontologies from the Linked Open Data cloud to explain a collection of tables. Firstly determine the Thematic Framework or the dominant concept(s) that the tables are collectively about and secondly, determine the Schematic Framework or entities/properties that each of the row values and column headers relate to in the context determined by the Theme. Such matched ontologies can be enriched in two ways: (a). they can be augmented with inference rules and new assertions to enable semantic reasoning within them, and (b). they can be interrelated to one another to form a global frame, that can support reasoning *across* them.

2 Related Work

Determining a meaningful context, extracting relevant ontologies and generating enriched RDF tuples from structured, unstructured and semi-structured data using appropriate ontologies from LOD cloud are all active research areas [6], [5], [7], [8].

We divide this broad literature into the following groups :

– **Identifying and Relating Concepts and Entities from Content**
Tools such as Open Calais[1], FRED [2], Apache Stanbol[2], Fox[3] work on unstructured content, extract concepts and entities such as places, events, people, organisations etc and relate them to universally known entities from knowledge bases such as DBPedia, Freebase, Geonames etc. While Open Calais and FRED amongst these are the most advanced tools with capabilities to extract context and related entities, the ontology/metadata they use internally are proprietary, in the sense that the disambiguated entities refer to an internal Calais or a FRED URI/id. Our objective is to extract concepts for the identified context that can be related to an openly available knowledge base from the Linked Open Data Cloud without using any proprietary vocabulary. In the context of datasets in LOD cloud, Lalithsena et

[1] Open Calais - http://viewer.opencalais.com/
[2] Apache Stanbol - https://stanbol.apache.org/overview.html
[3] FOX: Federated knOwledge eXtraction Framework - http://aksw.org/Projects/FOX.html

al. in [9] use an interesting technique to identify domains for such datasets with an aim to annotate/categorize the datasets appropriately. They rely on the Freebase knowledge base to identify topic domains for LOD.

- **Extraction of RDF tuples from CSVs**
  Several research efforts have addressed extraction of RDF tuples from CSV files. Some prominent tools in this area include RDF Converter[4], Virtuoso Sponger[5], Open Refine[6] and RDF123 [3]. However, the generated RDF tuples are mostly still raw data without any contextual reference. These RDF tuples need to be semantically linked to knowledge sources such as the ones constituting the Linked Open Data Cloud[7] or other formal ontologies to extract meaningful inferences from the data.
- **State of Art : Understanding Semantics of Tables and generating enriched RDF**
  Some of the most recent and relevant work that compares to our research includes work by Mulwad [4]. Mulwad's work is quite comprehensive in determining the meaning of a table and uses parameterized graphical model to represent a table. Their core module performs joint inferencing over row values, column headers and relations between columns to infer the meaning of the table by using a semantic message passing scheme that incorporates semantics into the messages. The graph is parameterised on three variables 1) one to determine the classes the column should map to 2) second to determine the appropriate relations between the data values in a row 3) third to determine relation between the column headers. The joint inferencing module is an iterative algorithm that converges when the model variables agree on the alloted LOD classes/entities for the column headers, relations between columns and row values. They also generate enriched RDF encapsulating the meaning of the table.

While these efforts are attractive and generate quality linked data keeping the intended meaning of the data in mind, they still work on a single table and are largely data values driven. In our challenge, we are looking for the ontology/collection of classes from related ontologies that fit best a *set* of tables, each table contributing a set of its columns to the identified ontology(ies). We expect a set of tables to have different utilitarian views depending upon the desired context. Our research aims to provide this semantic framework wherein a set of tables are mapped to domain from LOD. The columns from various tables are linked to relevant properties of the domain classes that the tables are collectively about, the data values and relations between columns in the tables are instantiations of the domain classes and their properties. Our overall objectives from this research is also, given a set of tables, generate enriched RDF data that can be further exploited by semantic reasoners.

---

4 RDFConverter: http://www.w3.org/wiki/ConverterToRdf

5 Virtuoso Sponger: http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VirtSponger

6 Open Refine: https://github.com/OpenRefine

7 Linked Data Design Issues: http://www.w3.org/DesignIssues/LinkedData.html

3 Research Questions and Hypothesis

My thesis attempts to answer the overall research objective to semantically integrate a collection of tabular data sets by inferring a Thematic and Schematic Framework. The following research questions detail the research objective.

1. Given a Collection of arbitrary tabular datasets, is it possible to determine what the collections of tables is about? Can we relate this inferred theme in terms of known concept(s) from Linked Open Data (LOD) or a custom knowledge base? We call this the Thematic Framework. We envisage the Thematic Framework to contain *Concept/Domain Classes* from LOD that best describe the collection of datasets.
2. For the identified Thematic Framework, can we relate the column headers in the various tables as properties of the identified concept classes from LOD and data values as instances or entities of the concept classes? We call this the Schematic Framework consisting of the T-Box (class and property definitions) and A-Box (class and property assertions). We envisage the Schematic Framework to contain a) properties of the dominant classes that are most relevant for the columns in the datasets in line with the identified Thematic Framework b) A-Box instantiations for all the data values in the datasets using the dominant classes and their properties c) T-Box definitions for the hierarchy of dominant classes and their properties derived from the respective vocabularies in LOD
3. Finally, can the enriched RDF generated using the Thematic and Schematic Framework discussed above, be processed by a reasoner such as Apache Jena to perform semantic inferences?

We hypothesise that for tabular datasets with data values that can be linked to LOD or some available custom knowledge base, it is possible to infer dominant concepts that relate to concept classes from known ontologies and further map the column headers and data values of the tables to properties and entities of those concept classes respectively. The definitions of properties and classes described explicitly in an ontology (DBPedia, Yago and others from Linked Open Data Cloud) and those implicitly derived from the instance assertions together with additional evidence from column headers can be combined with graphical modelling techniques to achieve the research objective. Table 1 shows a sample Thematic and Schematic Framework output for a set of two input data files (Table a , Table b) that have information on some Indian states and their capitals and rivers and their state of origin.

4 Proposal

Our first goal is to find dominant concept classes that relate maximally to a given set of tables. This paper showcases preliminary results towards this first goal. We propose two approaches for the concept class(es) identification and combine the two to obtain an overall scoring. In the *bottom-up* approach, entities are searched from LOD to obtain classes that maximally subsume the data values in a column. We also use the *bottom-up* technique to mine properties for the columns that best relate to the relation between the data values in a pair of columns. In the *top-down* approach, we rely completely on the

Table 1: Sample Thematic and Schematic Framework

Table a: StatesandCapitals.csv

State	Capital
Andhra Pradesh	Hyderabad
Maharashtra	Mumbai
Karnataka	Bangalore
Tamil Nadu	Chennai
Uttarakhand	Dehradun

Table b: RiversandSourceState.csv

River	Source
Ganges	Uttarakhand
Yamuna	Uttarakhand
Godavari	Maharashtra
Krishna	Maharashtra
Kaveri	Karnataka

Thematic Framework : Dominant Concept Classes
http://dbpedia.org/ontology/PopulatedPlace
http://dbpedia.org/ontology/River
Schematic Framework
StateandCapitals/State a dbpedia-owl:PopulatedPlace
StateandCapitals/Capital a dbpedia-owl:PopulatedPlace
RiversandSourceState/River a dbpedia-owl:River
RiversandSourceState/Source a dbpedia-owl:PopulatedPlace
StateandCapitals/Capital a owl:ObjectProperty
StateandCapitals/Capital rdfs:domain dbpedia-owl:PopulatedPlace
dbpedia.org/resource/Karnataka a dbpedia-owl:PopulatedPlace
dbpedia.org/resource/Bangalore a dbpedia-owl:PopulatedPlace
dbpedia.org/resource/Karnataka dbpedia-owl:Capital dbpedia.org/resource/Bangalore

column header literals or other information in the table description to arrive at candidate properties and their respective domain classes. For columns containing arbitrary literals, only top-down technique is applicable. We assume that the literals used to label the column headers are relevant to the data contained in the respective columns as otherwise, it will be practically impossible to ascertain what the data is about even by humans. We combine results from the top-down and bottom-up techniques and create a consolidated graph linking columns from tables to their respective candidate classes (derived from *bottom-up technique*) using *cc* edge label (*cc* used to denote candiate class link), and candidate properties for the columns (derived from *top-down technique* and *bottom-up technique*) to their respective *domain classes* using *d* edge label (*d* used to denote link to a domain class). We use DBPedia to generate the preliminary list of domain classes for the columns and call it the *Hypothesis Set* and expand the search for candidate properties to all the equivalent classes from LOD (determined by the *owl:equivalentClass* property for each domain class of the candidate property). This way we can identify dominant concept classes for a given set of tables across LOD. We use two Abduction Reasoning Heuristics namely a) Consistency and b) Minimality to arrive at the dominant class(es) [10].

Our Scoring Model to determine the dominant classes is as follows:

1. Candidate Class Support (CCS), defines how well a class $\gamma \in \Gamma$ fits as a candidate class for columns across all the CSV files:

$$ccs(\gamma) = \frac{\sum f_k}{cscols(\gamma)} \tag{1}$$

2. Domain Class Support (DCS) defines how well a class $\gamma \in \Gamma$ corresponds to candidate properties for columns across all the CSV files:

$$dcs(\gamma) = \frac{|dscols(\gamma)|}{|cols|} \tag{2}$$

Here, Γ represents the Hypothesis Set. Each member of this class γ will have incoming edges representing one of the following: a) candidate class for some column c_k with its corresponding support f_k represented by an incoming cc link, and/or b) domain class for some property p represented by an incoming d link. $\sum f_k$ is the sum of the support from each column connected to class γ with a cc link. $cscols(\gamma)$ is the set of nodes of type column that have a path leading to γ with a cc link. Similarly $dscols(\gamma)$ is the set of nodes of type column having a path to γ with a d link. $cols$ is the set of all nodes of type column.

ccs and dcs calibrate the prolific nature of the class across all the CSV files. In addition to the above scores, a "universality score" is associated with a class that describes how prolific is this class across different tables. This score called Tabular Support (TS) is defined as:

$$ts(\gamma) = \frac{|tabs(\gamma)|}{|tabs|} \tag{3}$$

Here, $tabs(\gamma)$ is the set of nodes labeled "table" in the graph that have a path to γ via any of the labeled edges and $tabs$ is the set of nodes labeled "table" in the graph.

The class score vector for class γ is a vector representing ccs and dcs scores:

$$csv(\gamma) = [ccs(\gamma), dcs(\gamma)]$$

The overall score representing the suitability of a class γ as a domain class is defined as:

$$Score(\gamma) = \|csv(\gamma)\|_2 \cdot H[csv(\gamma)] \cdot ts(\gamma) \tag{4}$$

Here $\|csv(\gamma)\|_2$ represents the L_2 norm of the class score vector and $H[csv(\gamma)]$ represents the entropy of the class score vector, given by:

$$H[csv(\gamma)] = -\sum_{i \in ccs(\gamma), dcs(\gamma)} p_i(\gamma) \log p_i(\gamma) \tag{5}$$

From the Overall Scores for each entry in the Hypothesis set, we use a *user defined threshold* to select the dominant concept(s) for the collection of tables. Our approach uses data-driven techniques and additional evidence from column headers and looks for convergence to domain classes in the context determined by the data. This is one of the differences from the State of the Art techniques discussed in *section 2*

5 Preliminary Results

As of this writing, we have considered a variety of tabular datasets ranging from hand crafted to publicly available csv datasets including those from data.gov.in.

Table 2 shows the preliminary results from our scoring model to identify dominant concept classes from a collection of tables using a cutoff threshold at 0.75.

Table 2: Dominant Concept Classes for the various collection of tabular datasets

Tabular Datasets	**Description**	**Dominant Concept Classes**
1) StatesCapitals.csv, 2) RiversSources.csv	Arbitrary Indian States and their capitals and Prominent Indian Rivers and their Source States	http://dbpedia.org/ontology/Place http://dbpedia.org/ontology/PopulatedPlace http://schema.org/Place http://schema.org/BodyOfWater http://dbpedia.org/ontology/BodyOfWater http://dbpedia.org/ontology/River http://schema.org/RiverBodyOfWater
1) PM.csv, 2) Prez.csv	Indian Prime Ministers and Presidents	http://xmlns.com/foaf/0.1/Person http://dbpedia.org/ontology/Person http://schema.org/Person
1) TechStartupUS.csv, 2) USStateCities.csv	Details of type and location of Technology Start-up companies in the US and arbitrary US State and Cities	http://dbpedia.org/ontology/Place http://schema.org/Place http://dbpedia.org/ontology/PopulatedPlace http://dbpedia.org/ontology/Location http://dbpedia.org/ontology/Organisation http://dbpedia.org/ontology/Settlement

6 Conclusion and Future Directions

The first goal of our research objective namely *Thematic Framework* extraction, as of now has been verifed with satisfactory results on tables, where the dominant concepts are about persons, places, organisations or some identifiable concept defined in LOD. The main challenge is the ability to identify LOD entities/resources from the data accurately especially when the Information Content/Entropy in the data from columns is low. Additionally the column header may capture the essence of the properties for a domain class using words that have a similar word-sense. We would like to test and refine the algorithm on variety of tables where the data values are a combination of known LOD entities and arbitrary values. Additionally our proposal faces challenges to converge to any dominant theme when the dataset is about a complex concept such as *Rice Prices on a particular date in various districts of India* or a dataset about *Real Estate Sales Transations in a particular locality*. Such instances occur when we do not have appropriate classes/ontologies in LOD that relate to the dataset in hand or the data values in the tables do not map to any entity in the LOD. In such cases, we would like to explore the use of SKOS(Simple Knowledge Organization System) categories and Yago concept classes together with Wordnet (to address the problem of similar words that capture the essence of the column header literals) as they seem to closely relate to the purpose/context of the data. Additionally, we would like to incorporate human input/custom ontology to validate the suggestions on concepts/properties returned by the algorithms. The next step from here is to expand the *Thematic Framework* to the corresponding *Schematic Framework* (T-Box and A-Box assertions) and device an appropriate scoring algorithm for abduced properties in line with the Thematic Framework and finally generate enriched RDF.

7 Evaluation Plan

We intend to use evaluation methods measuring a) Coverage b) Accuracy c) Applicability. **Coverage** will measure the percentage of tables/columns mapped to LOD classes.

The goal is to cover as many columns in all the datasets to relevant properties from LOD in line with the Thematic Framework abduced for the datasets. **Accuracy** will compare the scores of the Dominant Concept(s) in the Thematic Framework and scores of the properties suggested by the Schematic Framework with the actual concepts/properties suggested by human evaluators. **Applicability** will measure the relevance of the newly abduced A-Box instantiations and newly inferred LOD properties for relations between column headers and the Dominant Concepts for a collection of datasets.

Acknowledgments. I am deeply grateful to my thesis advisor Prof. Srinath Srinivasa for his support.

References

1. Srinivasa, Srinath and Agrawal, Sweety V. and Jog, Chinmay and Deshmukh, Jayati: Characterizing Utilitarian Aggregation of Open Knowledge In: Proceedings of the 1st IKDD Conference on Data Sciences, pp. 6:1–6:11. ACM, New York 2014
2. Presutti, Valentina, Francesco Draicchio, and Aldo Gangemi. Knowledge extraction based on discourse representation theory and linguistic frames. In Knowledge Engineering and Knowledge Management, pp. 114-129. Springer Berlin Heidelberg, 2012.
3. Han, Lushan, Tim Finin, Cynthia Parr, Joel Sachs, and Anupam Joshi. RDF123:from Spreadsheets to RDF. Springer Berlin Heidelberg, 2008.
4. Mulwad, Varish Vyankatesh. TABEL - A Domain Independent and Extensible Framework for Inferring the Semantics of Tables. PhD diss., University of Maryland, 2015.
5. Buitelaar, Paul, Philipp Cimiano, and Bernardo Magnini. Ontology learning from text: An overview. Vol. 123. 2005.
6. Lau, Raymond YK, Jin Xing Hao, Maolin Tang, and Xujuan Zhou. Towards context-sensitive domain ontology extraction. In System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on, pp. 60-60. IEEE, 2007.
7. Gerber, Daniel, Sebastian Hellmann, Lorenz Buhmann, Tommaso Soru, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. Real-time RDF extraction from unstructured data streams. In The Semantic Web - ISWC 2013, pp. 135-150. Springer Berlin Heidelberg, 2013.
8. Augenstein, Isabelle, Sebastian Pado, and Sebastian Rudolph. Lodifier: Generating linked data from unstructured text. In The Semantic Web: Research and Applications, pp. 210-224. Springer Berlin Heidelberg, 2012.
9. Lalithsena, Sarasi, Pascal Hitzler, Amit Sheth, and Paril Jain. Automatic domain identification for linked open data. In Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, vol. 1, pp. 205-212. IEEE, 2013.
10. Asha Subramanian, Srinath Srinivasa, Pavan Kumar, and S. Vignesh. 2015. Semantic Integration of Structured Data Powered by Linked Open Data. In Proceedings of the 5th International Conference on Web Intelligence, Mining and Semantics (WIMS '15) ACM, New York, NY, USA.

Multi-level context adaptation in the Web of Things

Mehdi Terdjimi

Université de Lyon, LIRIS
Université Lyon 1 - CNRS UMR5205
F-69622, France
mehdi.terdjimi@liris.cnrs.fr

Abstract. The Web of Things (WoT) aims at connecting things to applications using web technologies, on top of the Internet of Things. WoT applications are distributed and gather different levels of abstraction. They must be scalable and adapt to dynamic changes in their environment. The question we explore in the scope of my PhD thesis is: how can we deal with context in WoT applications? Our objective is to enable scalable multi-level context-aware adaptation for WoT applications. We intend to build models to describe context and reason about it. First, we have studied related work to identify a set of contextual levels and dimensions and have proposed semantic models suitable for several adaptation tasks in WoT applications. Second, we designed and implemented an architecture that distributes some adaptation tasks onto the client side, to improve reasoning scalability.

1 Introduction

The Internet of Things (IoT) aims at connecting devices (i.e. sensors and actuators) to the Internet, to share information using various protocols. The Web Of Things (WoT) builds upon the IoT, where connected devices ("things") rely on Web technologies and standards to break through silos and allow interoperability in pervasive applications [15], e.g. by making use of semantics.

In ubiquitous computing, context highly impacts application behavior and is composed of various pieces of information. Our research question is: how can we deal with context in WoT applications? And subsequently, how can we model context in a generic yet efficient way: on one hand, to harvest contextual data from different sources and build coherent and reliable models? On the other hand, how to actually perform the adaptation process for different purposes, required by WoT applications? To answer these questions, we propose in this thesis the concept of multi-level adaptation.

1.1 Context of the work: the ASAWoO project

In our work as part of the ASAWoO project[1], we propose the concept of avatar to augment and represent physical objects in the virtual world. The avatar of a

[1] http://liris.cnrs.fr/asawoo/

physical objects exposes its functionalities as RESTful resources. In some cases, some features of the object might be disabled for technical reasons such as physical constraints, or due to security, privacy, or other policies. The decision to make features available or not depends on context. To enable adaptation of connected devices to the changes affecting them, appropriate context modeling is required.

1.2 Challenges and motivations

In this work, we aim at using Web standards, such as service sharing protocols (through REST architectural style) and semantic Web technologies (RDF, OWL, SPARQL). The current challenges are:

1. To design generic, semantically-annotated context models for WoT applications, based on the state of the art, to allow context reasoning and adaptation. This will allow interoperability among heterogeneous contextual data sources, such as sensors of the object, external information gathered through Web services and application domain knowledge.
2. To enable multi-level adaptation, as WoT applications cover multiple abstractions, domains and needs. The adaptation would be realized by an appropriate engine, designed in two parts. The first part would be in charge of the extraction of relevant information (depending on the chosen adaptation type). The second part would be the adaptation engine itself.
3. To provide a scalable adaptation engine, in view of the increasing number of heterogeneous connected devices. A modularization of the reasoning steps would allow their distribution between the avatar and its clients. Each step could be executed on the client if its computing resources are sufficient.

In Section 2, we propose a state of the art on context modeling, followed by a state of the art on mobile/client reasoning. We propose in Section 3 a generic, flexible and scalable context model that constitutes our approach. A multi-level semantic adaptation process is presented and evaluated, with a method that locates the reasoning steps between the client and the WoT infrastructure in Section 4. In Section 5, we discuss the results and give the perspectives and future work.

2 Related work

In this section, we overview related work in the field of context modeling to help us build context models. We also study related work in the domain of mobile reasoning to help us build our adaptation solution.

2.1 Context modeling

Former definitions of context are generic and define several dimensions such as Location, Environment, Time, Activity, and User (Schilit and Theimer [26], Pascoe [21], Dey [10], Schmidt [27]). Context models used in IoT applications are

close to the former, but adapted to particular needs [22]. Some of them focus on physical aspects by using context dimensions related to the presence, the activity and the state of entities (i.e. people and devices) in some location, at a certain time [25, 11, 27, 35, 8, 1]. Some works use network context to provide efficient routing and disruption-tolerance [13, 18, 20, 32, 23]. In the field of social computing, context models have been designed for different purposes. To facilitate user interactions with the application [6, 33], or to improve organization within multi-agent systems [4, 3, 2, 5]. There are also works that separate application architecture information and business logic [16, 7, 19, 12], or the device and its physical properties [31, 30]. Another popular usage of context is related to content adaptation, which could be media [34] or, more recently, linked data [9].

The designed context models are specific to the application. But none of them completely rely on the Web, nor perform adaptation on multiple abstraction levels. Thus, as far as we know, there is no multi-level context model for WoT applications.

2.2 Mobile reasoning

When it comes to reason about context to perform adaptation, client-side reasoning may be a solution to address scalability concerns that arise with high numbers of simultaneous requests. But even if client processing resources augment at a fast pace, they remain heterogeneous and in some cases, too limited to execute heavy calculation processes. Thus, adaptivity and flexibility depending on the client's resources are necessary.

The following approaches aim at optimizing the reasoning process for resource-constrained devices. Different ways have been envisioned, from axiom-template rewriting (Kollia and Glimm [17]), to the Triple Pattern Fragments approach (TPF) which relies on intelligent clients that query TPF servers to address the problem of scalability and availability of SPARQL endpoints. However, the use of a server is always necessary. Concerning mobile reasoners, some are based on first-order logic (FOL) (KRHyper [28]), or description logics (DL) (Mine-ME 2.0 [24], $\mathcal{EL}+$ Embedded Reasoner [14]). But KRHyper is not designed for DL, and [24, 14] do not provide a web client access. An approach to embed a reasoner in mobile devices is to rely on web standards and run it in a web browser in Javascript. EYE[2] is a Node.js[3]-compatible reasoner, limited to FOL, which only runs on the server-side. Based on the JSW Toolkit, OWLReasoner[4] allows client-side processing of SPARQL queries on OWL 2 EL ontologies. As far as we know, it is the only full-Javascript OWL 2 EL reasoner that can be used offline in a web client, though its SPARQL engine is limited to basic rule assertions.

[2] http://reasoning.restdesc.org/
[3] https://nodejs.org/
[4] https://code.google.com/p/owlreasoner/

3 Approach

Our approach is based on the research questions raised in Section 1. First we plan to model each abstraction level based on the dimensions identified in the state of the art. As our context model will be described in an ontology, we consider an abstraction as a context state corresponding to a graph part. An example is depicted in Figure 1. In this example, the question *"Which communication protocols can be used?"* will be queried on the graph which contains the state `(Location: home, Security: Level_1, Time: Evening)`. In some cases, a dimension has no use in the abstraction (e.g. the Gender in the Communication level). Secondly, we plan to model the context of the WoT application state, in order to provide adaptation of the reasoning task. To do this, we evaluate our implementation in Section 4 to identify which parameters to model.

	Location	Security	Time	Gender
SOCIAL	Home	Level 1	Evening	Female
APP. ARCHITECTURE	Ø	Level 3	Ø	Ø
COMMUNICATION	Home	Level 1	Evening	Ø
PHYSICAL	Indoor	Level 1	Evening	Ø

Fig. 1. Example illustrating our approach

The overall method is: 1) to generate graphs from the context models, 2) to query these graphs in SPARQL, in order to retrieve each possibility given a context state, and 3) to add a rule engine that drives the reasoning process. Each step would be separated, to be deferred in the client side. Thus, we propose in the next Section an implementation to reach these goals.

4 Implementation and evaluation

We implemented an architecture that allows the modularization of the reasoning steps and client-side code migration. We proposed in [29] the HyLAR prototype[5] (for Hybrid Location-Agnostic Reasoning), which is a lightweight, modular and adaptive architecture developed in Javascript for hybrid client/server side reasoning, based on OWLReasoner. It allows client-side processing of SPARQL queries on OWL 2 EL ontologies, and consists in the separation of JSW modules that perform ontology classification (JSW Classifier), ontology and SPARQL query parsing (JSW Parser) and reasoning (JSW Reasoner)[6]. JSW modules are

[5] The prototype is available at http://dataconf.liris.cnrs.fr/owlReasoner/

[6] Originally, OWLReasoner rewrites the aBox into a relational database and SPARQL queries into SQL queries. This choice is not ours and will not be discussed here.

packaged as Node.js modules and AngularJS[7] services. This way, they can be executed on either the server or client. On the client side, the reasoner modules can be embedded either in a regular angular service, or in a web worker. As the main service is totally agnostic about the location of the reasoning modules, security and privacy concerns arise, but these aspects are subject to future work.

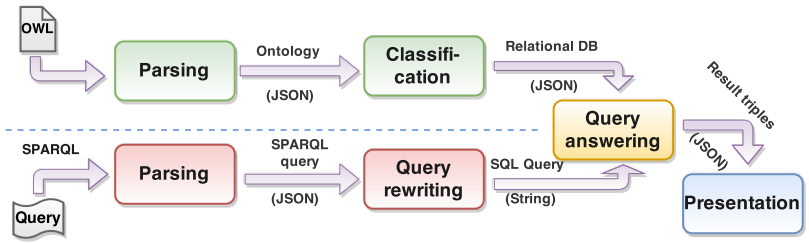

Fig. 2. OWLReasoner's reasoning steps

The goal of the evaluation we propose in [29] is to identify the parameters affecting the reasoning task's processing time, and to find an optimal configuration. We calculated the overall reasoning process request-response times in three situations: full server-side, full client-side and hybrid (server-side parsing and classification, and client-side query processing). For the latter two variants, client-side parts are evaluated both with and without web worker. We assume that scripts and ontologies are available on the server. Each scenario is evaluated with these initial parameters and tools:

- Ontology sizes: ontology A has 1801 class assertions and 924 object property assertions, and B has 12621 class and no object property assertions)[8]. These datasets are conference keywords gathered from DataConf[9].
- Network status: requesting locally or in high latency conditions (around 150ms, using Clumsy 0.2[10]).
- Client capabilities: a Dell Inspiron (i7-2670QM CPU @ 2.20GHz, which also hosts the Node.js server), a Nokia Lumia 1320 (Snapdragon S4 @ 1700 MHz) and a Samsung Galaxy Note (ARM cortex A9 Dual-Core @ 1,4 GHz).

In Tables 1 and 2, [Q] is the time for the client's request to reach the server, [P] is the processing time and [R] is the time for the server response to reach the client. Some parts of these steps/patterns are considered immediate and noted in the result tables as not applicable.

[7] http://www.angularjs.org
[8] Due to OWLReasoner query engine limitations that does not currently allow querying individuals nor data property assertions, our evaluations are limited to class and object property assertions.
[9] http://dataconf.liris.cnrs.fr/
[10] http://jagt.github.io/clumsy/

Ontologies A / B	**[R0]**	**[Q1]**	**[R1]**	**[Q2]**	**[R2]**	**[Q3]**	**[R3]**
Remote server	334	54	110 / 275	119 / 120	167 / 647	146 / 154	61 / 85

Table 1. Network delays (in ms)

Ontologies A / B	**[P2] (no worker)**	**[P2] (worker)**	**[P3] (no worker)**	**[P3] (worker)**
Inspiron (Chrome)	790 / 27612	764 / 26464	28 / 101	24 / 88
Lumia (IE)	1989 / 54702	1883 / 53801	156 / 198	144 / 185
Galaxy Note (Firefox)	2954 / 81255	2872 / 79752	465 / 2988	440 / 2872
Server (Node.js)	780 / 20972	n/a	35 / 37	n/a

Table 2. Classification [P2] and reasoning [P3] times (in ms)

5 Discussion and future directions

5.1 Analysis of evaluation results

As expected, Table 2 shows that the server has the best results for the classification processing time and can use caching. Even if the raw ontology is faster to load than the classification results, loading scripts and data on the client is much faster than performing the same classification step on each client. Thus, it makes no sense to migrate heavy calculations onto clients, rather than pre-calculating them on the server and caching results. More generally, for M clients and N queries/client, we calculate each configuration calculation times as follows[11]:

- Full server-side: $P2_{server} + M \times N \times (Q3 + P3_{server} + R3)$
  We group network (Q, R) and application (P) statuses as the full process is server-side.
- Full client-side: $M \times (R0 + Q1 + R1) + P2_{client} + N \times P3_{client}$
  Q, R and P fully depend on the client. They are therefore difficult to estimate in comparison to the full-server configuration.
- Hybrid: $P2_{server} + M \times (R0 + Q2 + R2) + N \times P3_{client}$
  Client P estimation is easier as it only concerns the query-answering process.

We identify the following parameters affecting the reasoning task: the number of clients (M) and queries per client (N), the network status (Q and R), and the ontology size and computing resources (P).

5.2 Lessons learned and future directions

There is no optimal configuration, and choosing a location for each step is a complex task as it is context-dependent. In our evaluation, we identified different parameters that affect processing times. They are crucial for the adaptation process and to respond to our research questions: how can we model the context 1) to allow the inclusion of the context elements needed by the application and 2) to perform the adaptation process for different purposes, specific to WoT applications? Our future work includes taking users' privacy into account, and

[11] Server-side classification (performed once and then cached) and client-side calculations (performed in parallel) are only counted once.

describing our context model in an ontology. Graphs corresponding to a particular context state would be generated and queried. We plan to reuse an existing adaptation engine that suits our needs for the reasoning process.

References

1. Arias, M.: Context-Based Personalization for Mobile Web Search. Context pp. 33–39 (2008)
2. Bazire, M., Brézillon, P.: Understanding context before using it. Modeling and using context (2005)
3. Brézillon, P., Pomerol, J.: Contextual knowledge sharing and cooperation in intelligent assistant systems. Le Travail Humain pp. 1–33 (1999)
4. Brézillon, P.: Context in Artificial Intelligence: II. Key elements of contexts. Computers and artificial intelligence pp. 1–27 (1999)
5. Bucur, O., Beaune, P., Boissier, O.: Representing context in an agent architecture for context-based decision making. Proceedings of the Workshop on ... (2005)
6. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click-through and session data. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '08 p. 875 (2008)
7. Chaari, T., Laforest, F., Flory, A., Einstein, A.A., Cedex, V.: Adaptation des applications au contexte en utilisant les services web. Proceedings of the 2nd French-speaking conference on Mobility and uibquity computing - UbiMob '05 (2005)
8. Coppola, P., Mea, V.D., Di Gaspero, L., Menegon, D., Mischis, D., Mizzaro, S., Scagnetto, I., Vassena, L.: The context-aware browser. IEEE Intelligent Systems 25(1), 38–47 (2010)
9. Costabello, L., Villata, S., Gandon, F.: Context-aware access control for rdf graph stores. In: ECAI. pp. 282–287 (2012)
10. Dey, A.K.: Understanding and using context. Personal and ubiquitous computing 5(1), 4–7 (2001)
11. Dey, A.K., Salber, D., Abowd, G.D., Futakawa, M.: The conference assistant: Combining context-awareness with wearable computing. In: Wearable Computers, 1999. Digest of Papers. The Third International Symposium on. pp. 21–28. IEEE (1999)
12. Gensel, J., Villanova-Oliver, M., Kirsch-Pinheiro, M.: Modèles de contexte pour l'adaptation à l'utilisateur dans des systèmes d'information web collaboratifs. In: Workshop from" 8èmes journées francophones". Sophia-Antipolis, France (2008)
13. Gold, R., Mascolo, C.: Use of context-awareness in mobile peer-to-peer networks. In: Distributed Computing Systems, 2001. FTDCS 2001. Proceedings. The Eighth IEEE Workshop on Future Trends of. pp. 142–147. IEEE (2001)
14. Grimm, S., Watzke, M., Hubauer, T., Cescolini, F.: Embedded $\mathcal{EL}+$ reasoning on programmable logic controllers. In: The Semantic Web–ISWC 2012, pp. 66–81. Springer (2012)
15. Guinard, D., Trifa, V., Mattern, F., Wilde, E.: From the internet of things to the web of things: Resource-oriented architecture and best practices. In: Architecting the Internet of Things, pp. 97–129. Springer (2011)
16. Kirsch-Pinheiro, M., Gensel, J., Martin, H.: Representing context for an adaptative awareness mechanism. In: Groupware: Design, Implementation, and Use, pp. 339–348. Springer (2004)

17. Kollia, I., Glimm, B.: Optimizing sparql query answering over owl ontologies. arXiv preprint arXiv:1402.0576 (2014)
18. Mascolo, C., Capra, L., Emmerich, W.: Mobile computing middleware. In: Advanced lectures on networking, pp. 20–58. Springer (2002)
19. Munnelly, J., Fritsch, S., Clarke, S.: An aspect-oriented approach to the modularisation of context. In: Pervasive Computing and Communications, 2007. PerCom'07. Fifth Annual IEEE International Conference on. pp. 114–124. IEEE (2007)
20. Musolesi, M., Mascolo, C.: Car: context-aware adaptive routing for delay-tolerant mobile networks. Mobile Computing, IEEE Transactions on 8(2), 246–260 (2009)
21. Pascoe, J.: Adding generic contextual capabilities to wearable computers. In: Wearable Computers, 1998. Digest of Papers. Second International Symposium on. pp. 92–99. IEEE (1998)
22. Perera, C., Zaslavsky, A., Christen, P., Georgakopoulos, D.: Context aware computing for the internet of things: A survey. Communications Surveys & Tutorials, IEEE 16(1), 414–454 (2014)
23. Raverdy, P.G., Riva, O., de La Chapelle, A., Chibout, R., Issarny, V.: Efficient context-aware service discovery in multi-protocol pervasive environments. In: Mobile Data Management, 2006. MDM 2006. 7th International Conference on. pp. 3–3. IEEE (2006)
24. Ruta, M., Scioscia, F., Loseto, G., Gramegna, F., Ieva, S., Di Sciascio, E.: Minime 2.0: powering the semantic web of things. In: 3rd OWL Reasoner Evaluation Workshop (ORE 2014)(jul 2014) (2014)
25. Schilit, B.N., Adams, N., Gold, R., Tso, M.M., Want, R.: The parctab mobile computing system. In: Workstation Operating Systems, 1993. Proceedings., Fourth Workshop on. pp. 34–39. IEEE (1993)
26. Schilit, B.N., Theimer, M.M.: Disseminating active map information to mobile hosts. Network, IEEE 8(5), 22–32 (1994)
27. Schmidt, A.: Ubiquitous computing-computing in context. Ph.D. thesis, Lancaster University (2003)
28. Sinner, A., Kleemann, T.: Krhyper–in your pocket. In: Automated Deduction–CADE-20, pp. 452–457. Springer (2005)
29. Terdjimi, M., Médini, L., Mrissa, M.: HyLAR: Hybrid Location-Agnostic Reasoning. In: ESWC Developers Workshop 2015. pp. 1–6. Portoroz, Slovenia (May 2015)
30. Truong, H.L., Dustdar, S., Baggio, D., Corlosquet, S., Dorn, C., Giuliani, G., Gombotz, R., Hong, Y., Kendal, P., Melchiorre, C., et al.: Incontext: A pervasive and collaborative working environment for emerging team forms. In: Applications and the Internet, 2008. SAINT 2008. International Symposium on. IEEE (2008)
31. Truong, H.L., Juszczyk, L., Manzoor, A., Dustdar, S.: ESCAPE–an adaptive framework for managing and providing context information in emergency situations. Springer (2007)
32. Wei, Q., Farkas, K., Prehofer, C., Mendes, P., Plattner, B.: Context-aware handover using active network technology. Computer Networks 50(15), 2855–2872 (2006)
33. Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E., Li, H.: Context-aware ranking in web search. Sigir 2010 p. 451 (2010)
34. Yu, Z., Zhou, X., Zhang, D., Chin, C.Y., Wang, X., et al.: Supporting context-aware media recommendations for smart phones. Pervasive Computing, IEEE 5(3), 68–75 (2006)
35. Zimmermann, A., Lorenz, A., Oppermann, R.: An operational definition of context. In: Modeling and using context, pp. 558–571. Springer (2007)

Efficient and Expressive Stream Reasoning with Object-Oriented Complex Event Processing

Riccardo Tommasini

Politecnico di Milano
riccardo.tommasini@polimi.it

Abstract. RDF Stream Processing (RSP) engines - systems able to continuously answer queries upon semantically annotated information flows - empirically proved that Stream Reasoning (SR) is feasible. However, existing RSP engines do not investigate the trade-off between the reasoning expressiveness and the performance typical of information flow processing (IFP) systems: either an high throughputs with a low expressiveness (e.g. ρDF) or an high expressiveness (e.g., $\mathcal{EL}$) with a low throughputs are provided. *Can the systematic exploration of this trade-off lead SR to continuously execute expressive reasoning without loosing the efficiency typical of IFP systems?* In this paper, we propose a Systematic Comparative Research Approach (SCRA) to investigate the RSP solution space. Moreover, in contrast with the state-of-the-art trend of adding IFP capabilities to reasoners, we discuss how to realize an Efficient and Expressive Stream Reasoning by adding reasoning capabilities into IFP systems (in particular to Object-Oriented Complex Event Processors).

1 Scene Setting

Stream Reasoning (SR) is a novel research trend that aims at enabling reasoning on rapidly changing information flows [12]. So far, many RDF Stream Processing (RSP) engines - systems able to cope with semantically annotated data flows - were developed as proof-of-concepts [1, 5, 6, 8]. However, due to the complexity of the reasoning task strongly impacts real-time processing, existing solutions either focus on keeping either Information Flow Processing (IFP [11]) comparable performances [16] offering low expressive reasoning (i.e., ρDF [18] with limited extensions) or to optimize expressive reasoning algorithms to the streaming scenario loosing the typical IFP performances [20].

The most of the state-of-the-art solutions pipeline IFP and Semantic Web reasoning modules into *black box* (BB) architectures.

White box (WB) architectural approaches, which redesign all the underlying modules into an integrated solution, can better investigate the performances and reasoning expressiveness trade-off [22]. WB attempts like [1, 7, 16] try to add IFP capabilities to reasoners, while the opposite approach, adding reasoning capabilities to IFP systems, is not attempted yet.

RESEARCH QUESTION: *Can the systematic exploration of the performances and reasoning expressiveness trade-off lead SR to continuously execute expressive reasoning without loosing the efficiency typical of IFP systems?*

Exploring the RSP solutions space requires to analyze the RSP engine while is processing. But, due to the complexity of the RSP engine, it might be hard. We need to enable a systematic comparative research approach (SCRA) [9] that simplifies the analysis through a strategy for cross-case studies. Moreover, realizing an efficient and expressing stream reasoning (E_2SR) demands both to rethink the execution semantic model and to expand the solution space with new implementations. Our approach aim at adding reasoning capabilities into IFP systems, in particular Object Orient Complex Event Processor.

Outline - the remainder of this paper is organized as follows: Section 2 summarizes state-of-the-art RSP engines with an IFP background and Section 3 presents a brief overview on RSP Benchmarking. Section 4 presents the proposed research approach. Section 5 describes the approach implementation and the current stage of development. Section 6 shows the evaluation methodology and summarizes the obtained results. Section 7 comes to conclusion presenting the work already done, the lessons learned, and our future directions.

2 RSP Engines State of the Art

Semantic Web (SW) and IFP technologies like Data Stream Management Systems (DSMS) or Complex Event Processing (CEP) played a crucial role to demonstrate that SR is possible. Indeed, they foster the definition of SR requirements [17]: (R.1) Real-Time processing (DSMS); (R.2) pattern-matching on incoming information (CEP); (R.3) Data Integration (SW); (R.4) Rich Ontology Languages (SW). And finally, (R.5) expressive query languages and (R.6) systems scalability (IFP & SW).

Table 1 summarizes and extends a recent survey [17]. It highlights some relevant characteristics of the state-of-the-art RSP engines with a IFP background:

- *Continuous Query Answering* [3] - it is needed to satisfy (R.1);
- *Background Data* - supporting static data access is needed to satisfy (R.3);
- *Time Model* - the system temporal model: one or more timestamps (R.1);
- *Reasoning* - the reasoning expressiveness, when reasoning is available (R.4);
- *Time-Aware* - time-related operators are crucial to satisfy (R.2);
- *Data Transformation* - presence of abstraction functions/aggregates (R.5);
- *Historical Data* - availability of historical data storages (R.3);
- *System Design* - w.r.t IFP or SW: DSMS/CEP/rule-based (R.6);
- *Architectural Approach* - the adopted architectural approach: white box (WB) or a black box (BB) (R.6).

RSP engines like Streaming Knowledge Base [24] and C-SPARQL Engine [5] are examples of in-memory, window-based, RSP engines that adopt a black box approach pipelining a DSMS and a naïve reasoner. They allow continuous query answering on RDF streams or graphs w.r.t background knowledge by the means

System	Cfr.	Cont. Queries	BG Data	Time Model	Reasoning	Time. Aware	Data Trans	Hist. Data	Arch. Appr.	Design
C-SPARQL E.	[5]	✓(p)	✓	TS	RIF*	✓***	✓		BB	DSMS
IMaRS	[4]	✓(p)	✓	TS	Transitive		✓		BB	DSMS
TrOWL	[20]	✓	✓	TS	EL+/SHIQ**				WB	Rules
CQELS	[7]	✓	✓	TS			✓		WB	DSMS
SKB	[24]	✓	✓	TS	OWL sub				BB	DSMS
SparkWave	[16]	✓	✓	TS	RDFS subset		✓		WB	Rules
ETALIS	[1]	✓	✓	2xTS	RDFS subset	✓	✓	✓	BB	CEP
Morph$_{stream}$	[7]	X	✓	TS	ELIO				BB	DSMS

Table 1. State Of The Art of RSP engines related to IFP - p:(periodic) WB: white box BB: black box; 2TS:interval; *:supports Jena Rule-Based Reasoning; **TBox/ABox; *** Subset of Allen Algebra by TS function

of queries expressed with extensions of SPARQL 1.1 (e.g. C-SPARQL) that include the time semantics.

Morph$_{stream}$ [7] ports some reasoning capabilities into existing DSMS system and allows to query virtual RDF streams with SPARQL$_{stream}$. A conjunctive query is translates into the union of multiple conjunctive queries thanks to a reasoner that performs query rewriting and an R2RML mappings extension with time semantics (windows). Queries are executed by an underlying DSMS.

ETALIS [2] engine is a WB solution that processes queries written in ETALIS languages converting them to Prolog rules and executes them on a Prolog engine at run-time. EP-SPARQL [1] is a SPARQL extensions for Event Processing that enables black box Stream Reasoning on ETALIS [2]. EP-SPARQL queries are translated in logic expressions of the ETALIS Language. [1] is the only solution that allows to write complex patterns with time constraints on incoming events, that provides streaming and historical data integration and that is natively time-aware. However, its performance are not satisfying at all [19].

CQELS [8] implements a WB approach porting DSMS concepts (e.g. physical operators, data structures and query executor) into an SPARQL engine with no reasoning capabilities. It can operate queries optimization, because each phase of the processing is available.

SparkWave [16] is a white box ruled-based RSP engine designed for high RDFS performance reasoning over RDF Streams by extending RETE, a reasoning system algorithm, to process incoming information flows. IMaRS [4] optimizes incremental reasoning by relying on a fixed time window to predict expiration times. TrOWL [20] is an engine for efficient incremental ontology maintenance when updates are frequent (but not streaming). It does not rely on fixed time windows to predict the expiration time of streaming information, but it reduces reasoning complexity exploiting syntactic approximation. Despite big performance limitations, TrOWL is still relevant for our research. Indeed, it supports TBox stream reasoning of EL+ and approximate ABox stream reasoning of SHIQ expressiveness.

Notice that ASP-based solutions are out of the scope of this research because their reasoning capabilities and performances are non-comparable with the systems in the solution space we target to investigate.

3 RSP Benchmarking State of the Art

The SR community focuses on RSP engine evaluation. So far, challenges and requirements were formulated [21] and many attempts to address them were developed [14, 19, 25]. Preliminary evaluations on the state-of-the-art confirmed that none of existing RSP engines provides IFP-comparable performances and expressive reasoning at the same time. However, RSP benchmarking still presents some limitations and it is not applied systematically yet.

[25, 14, 19] neither face all the challenges nor satisfy all the requirements proposed in [21]. They provide ontologies, datasets and queries for the evaluation. The metrics set comprises query language coverage, throughput and recently [14] query results mismatch and correctness, but does not consider memory consumption and query execution latency. A minimal testing facility is provided by [19, 14], but without a method to lead the investigation.

The stage of analysis is also limited. Indeed, it consists into an average result after a predefined testing period, while the dynamics of the RSP engine during the entire test is not considered.

RSP benchmarking still misses both an infrastructure to design and test RSP engines performances and a methodology to investigate systematically the trade off. In summary, a SCRA that allows to design and execute comparable, reproducible and repeatable experiments in a controlled environment and, thus, provide a picture of the solution space.

4 Proposed Approach

In Section 3 we stated that both the black box (BB) and white box (WB) state-of-the-art RSP engines show performance limitations [25, 14, 19]. The former cannot perform cross-module optimization, while the latter is realized by adding IFP-capabilities to reasoning systems and, thus, it is not possible to exploit the typical order-based optimization that guarantee IFP performances.

Building on these lessons learned, it would be possible to develop an efficient and expressive SR (E_2SR). Indeed, a common step to all the mature research areas is focusing on improving the systems performances [15]. Two approaches are possible, eliminating the lacks or reinventing the technology pillars. Both require to enable a systematic comparative research approach (SCRA) for RSP engines. Why should the investigation be comparative? SCRA is popular in those research fields where the complexity of the subject goes beyond the possible observable models (e.g. social science). Single-case studies help to deeply understand the subject, but do not foster any generalization. On the other hand, cross-case studies allow general thinking, but with and high final complexity.

Comparing RSP engine dynamics during the entire test execution will clarify how the actual execution semantic of the RSP engine influences the performances. SR needs a strategy to reduce the analysis complexity without losing the relevance of each involved system. SCRA consists into comparing RSP engine dynamics under a given experimental condition.

Enabling SCRA is crucial at this stage of development. A specific investigation methodology is required to contrast the performance measurements, state which solution is better, if any, and possibly drill down or raise up the analysis at different levels. We need to describe normality and stressing conditions for RSP engines, so it is necessary to understand which variables involve into the evaluation. Finally, it would be possible to investigate how the system actually works and to position it in the solution space. To this final extent, we need an infrastructure to design and systematically execute repeatable and reproducible experiments on a given RSP engine under comparable conditions.

The E_2SR goal are the high entailment regime of [20], (i.e. $\mathcal{EL}+$) and the IFP-compatible performances performances of [16, 7]. This requires a WB approach for intra-modules optimization. In Section 2 we presented the limitations of porting IFP-capabilities into reasoners. Moreover, [1] already covers all the featured characteristic that Table 1 highlights, but it is not optimized neither in performance nor for reasoning expressiveness. Thus, we propose to introduce reasoning capabilities in IFP systems, by extending rule-based Object Orient Complex Event Processor engines into a WB approach.

[1] - and in general CEP-based RSP engines - presents many technical opportunities towards E_2SR: (i) event processing languages like Tesla [10] or EPL[1] can perform the reasoning tasks that can be encoded as rules; (ii) they are natively order-aware, more specifically time-aware since data are ordered by recency [13, 22]. (iii) the information flows are usually represented with objects, as in object-oriented languages or databases. Object-Oriented programming languages natively allow some reasoning task which can improve the final entailment regime. (iv) last but not least, CEP systems are usually well-engineered, because the IFP research focused on bandwidth and latency performance optimization as well as scaling by the means of system distribution.

Finally, how to evaluate the obtained results? E_2SR performance evaluation explicitly needs to enable SCRA; SCRA itself demands instead to prove that: (i) the testing infrastructure does not influence the systems results; (ii) a base measurements-set and an investigation method are defined and accepted by the SR community; (iii) some baselines and analysis guidelines are available.

5 Approach Implementation

The first research phase focused on the following research sub-questions:

– SP.1 *Can a test-stand*[2] *enable a SCRA for RSP engines?*

[1] http://bit.ly/1GdhUFC

[2] an aerospace engineering facility to design and execute experiments over engines and to collect performance measurements

– SP.2 *It is possible to implement simple, ruled-based reasoning upon a CEP?*

SP.1) My Master Thesis had the goal to develop Heaven [23], an open source[3] framework that consists into an RSP engine test stand, four naïve implementations of black box DSMS-based RSP engines called baselines and a evaluation methodology. Heaven targets window-based, in-memory RSP engines implemented in Java, like C-SPARQL engine [5] or CQELS [8] or the baselines themselves. The test-stand makes no assumption about the tested RSP engine internal process, treating it as a black box. Thanks to Heaven, it is finally possible to design comparable, repeatable and reproducible experiments by providing: an RSP engine $\mathcal{E}$, an ontology $\mathcal{T}$, a query-set $\mathcal{Q}$ and an dataset $\mathcal{D}$ to stream.

SP.2) Table 2 summarizes the current stage of development. Some E_2RS prototypes were implemented[4] with Esper, an open source CEP engine popular in the SR research field[5].

System	EPL	BG Data	Data Model	Sound/ Complete*
PLAIN	YES	Hashtable	Serialised	Yes
OO-STD	YES	OO	OO-RDF	Yes/No
GENERICS	YES	OO	OO-RDF	Yes/No

Table 2. E_2RS Prototypes * w.r.t ρDF [18]

The E_2RS prototypes development relies on the following assumptions, inspired by [16]: (i) the entailment regime is ρDF [18]; (ii) the initially ontology is small *static* and (iii) its materialization happend in pre-processing.

PLAIN encodes in EPL the rules for continuous query answer under ρDF entailment regime; events are serialized triples and the TBox is materialized within an hash-map. OO-STD proposes a solution where both EPL rules and Java polymorphism are equally exploited to perform typical reasoning tasks of ρDF (i.e., class hierarchy subsumption). Finally, GENERICS tries to extend OO-STD exploiting Java-generics.

6 Empirical Evaluation

SCRA and E_2SR evaluations are related. The former is evaluated by proving the effectivenesses of the testing infrastructure and of the investigation methodology. The latter requires to compare new performance results with the state-of-the-art solutions. Thus, we consider to start with SCRA evaluation, because E_2SR one strictly relies on it.

To demonstrate Heaven effectiveness we run some experiments on the test stand, which results are available[6]. As RSP engine we used some baselines, four window-based RSP engine implementations that are realized pipelining Esper

[3] https://github.com/streamreasoning/heaven

[4] https://github.com/streamreasoning/Proto-EESR

[5] Esper is written in Java and it supports sliding-windows (e.g [5] uses it to implement a black box RSP engine). Moreover, it exploits the EPL query language that has all the characteristics that we need (Section 4).

[6] http://streamreasoning.org/TR/2015/Heaven/iswc2015-appendix.pdf

and Jena ARQ. The baselines offer both RDFS naïve reasoning, materialization of the entire content of the active window at each cycle, and the incremental reasoning, maintaining the materialization over time by updating the differences between two consecutive windows. Our experiments empirically proved that Heaven influences on systems performance are stable and predictable. Thus, we can assert that it enables SCRA for RSP engine.

From the obtained insight, what is already clear is that even when an RSP engines is extremely simple (e.g., one of the baselines), hypothesis verification is hard (e.g. we cannot confirm that the incremental reasoning baselines outperform those with a naïve reasoning approach [20]).

7 Conclusion

The impact of enabling a systematic comparative research approach for RSP engine is potentially high in the Stream Reasoning research field. The initial insights we got by evaluating the baselines with Heaven showed how less we know about the RSP engine dynamics.

SCRA is a priority for all the SR community. Our first future work consist in systematically testing all the supported RSP engines, i.e only in-memory, window-based RSP engines developed in Java. To realize this we need to (i) implement an adapting facade for the RSP engines to test and (ii) define a suite of experiments, which exploits existing RDF streams, ontologies and queries available in the the state-of-the-art of RSP benchmarking [19, 25, 14] to show any aspects of the RSP engine dynamics.

About E_2SR, Esper-based prototypes (Table 2) are promising, but surely far from our goals (i.e. the [20] expressiveness and the [16, 7] performances). E_2SR research priority is defining a standard for event processing query language. E_2SR prototypes exploit a specific one: EPL. The future works should consider: (i) the definition of a minimal fragment of EPL to enable E_2SR; (ii) the prototyping of systems on alternative query languages like Tesla [10]; (iii) the proposal of a execution semantics for SR system built on CEP.

Acknowledgments. Thanks to my advisor Prof. Emanuele Della Valle (Politecnico di Milano) and my co-advisors Daniele Dell'Aglio and Marco Balduini.

References

1. Anicic, D., Fodor, P., Rudolph, S., Stojanovic, N.: EP-SPARQL: a unified language for event processing and stream reasoning. In: WWW 2011. pp. 635–644 (2011)
2. Anicic, D., Rudolph, S., Fodor, P., Stojanovic, N.: Stream reasoning and complex event processing in ETALIS. Semantic Web 3(4), 397–407 (2012)
3. Arasu, A., Babu, S., Widom, J.: The CQL continuous query language: semantic foundations and query execution. VLDB J. 15(2), 121–142 (2006)
4. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Incremental reasoning on streams and rich background knowledge. In: ESWC 2010. pp. 1–15 (2010)

5. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Querying RDF streams with C-SPARQL. SIGMOD Record 39(1), 20–26 (2010)
6. Bolles, A., Grawunder, M., Jacobi, J.: Streaming SPARQL - extending SPARQL to process data streams. In: The Semantic Web: Research and Applications, pp. 448–462. Springer Berlin Heidelberg (2008)
7. Calbimonte, J., Corcho, Ó., Gray, A.J.G.: Enabling ontology-based access to streaming data sources. In: The Semantic Web - ISWC 2010. pp. 96–111 (2010)
8. Calbimonte, J., Jeung, H., Corcho, Ó., Aberer, K.: Enabling query technologies for the semantic sensor web. Int. J. Semantic Web Inf. Syst. 8(1), 43–63 (2012)
9. Creswell, J.W.: Research Design: Qualitative, Quantitative, and Mixed Methods Approaches. Sage Publications Ltd., 3 edn. (2008)
10. Cugola, G., Margara, A.: Tesla: A formally defined event specification language. In: ACM International Conference on DEBS. pp. 50–61. ACM (2010)
11. Cugola, G., Margara, A.: Processing flows of information: From data stream to complex event processing. ACM Comput. Surv. 44(3), 15 (2012)
12. Della Valle, E., Ceri, S., van Harmelen, F., Fensel, D.: It's a streaming world! reasoning upon rapidly changing information. IEEE Intelligent Systems (2009)
13. Della Valle, E., Schlobach, S., Krötzsch, M., Bozzon, A., Ceri, S., Horrocks, I.: Order matters! harnessing a world of orderings for reasoning over massive data. Semantic Web 4(2), 219–231 (2013)
14. Dell'Aglio, D., Calbimonte, J., Balduini, M., Corcho, Ó., Della Valle, E.: On correctness in RDF stream processor benchmarking. In: The Semantic Web - ISWC 2013. pp. 326–342 (2013)
15. Gray, J. (ed.): The Benchmark Handbook for Database and Transaction Systems (2nd Edition). Morgan Kaufmann (1993)
16. Komazec, S., Cerri, D., Fensel, D.: Sparkwave: continuous schema-enhanced pattern matching over RDF data streams. In: 6th ACM International Conference on Distributed Event-Based Systems, DEBS 2012. pp. 58–68 (2012)
17. Margara, A., Urbani, J., van Harmelen, F., Bal, H.E.: Streaming the web: Reasoning over dynamic data. J. Web Sem. 25, 24–44 (2014)
18. Muoz, S., Prez, J., Gutierrez, C.: Minimal deductive systems for RDF. In: Springer-Verlag. pp. 53–67. ESWC '07, Springer-Verlag, Berlin, Heidelberg (2007)
19. Phuoc, D.L., Dao-Tran, M., Pham, M., Boncz, P.A., Eiter, T., Fink, M.: Linked stream data processing engines: Facts and figures. In: The Semantic Web - ISWC 2012. pp. 300–312 (2012)
20. Ren, Y., Pan, J.Z., Zhao, Y.: Ontological stream reasoning via syntactic approximation. In: Proceedings of the 4th International Workshop on Ontology Dynamics (IWOD 2010). vol. 651. Citeseer (2010)
21. Scharrenbach, T., Urbani, J., Margara, A., Della Valle, E., Bernstein, A.: Seven commandments for benchmarking semantic flow processing systems. In: The Semantic Web - ESWC 2013. pp. 305–319 (2013)
22. Stuckenschmidt, H., Ceri, S., Della Valle, E., van Harmelen, F.: Towards expressive stream reasoning. In: Semantic Challenges in Sensor Networks (2010)
23. Tommasini, R., Della Valle, E., Balduini, M., Dell'Aglio, D.: Heaven test stand: towards comparative research on RSP engines. In: OrdRing 2015-5rd International Workshop on Ordering and Reasoning. p. 6p (2015)
24. Walavalkar, O., Joshi, A., Finin, T., Yesha, Y.: Streaming knowledge bases. In: International Workshop on Scalable Semantic Web Knowledge Base Systems (2008)
25. Zhang, Y., Pham, M., Corcho, Ó., Calbimonte, J.: Srbench: A streaming RDF/SPARQL benchmark. In: The Semantic Web - ISWC 2012. pp. 641–657