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 backend 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)