published as: Waitelonis, J.; Knuth, M.; Wolf, L.; Hercher, J. & Sack: H.: The Path is the Destination -- Enabling a New Search Paradigm with Linked Data. In Proceedings of the Workshop on Linked Data in the Future Internet at the Future Internet Assembly, December 16-17, 2010, Ghent, Belgium, CEUR Workshop Proceedings Vol. 700, 2010. (ISSN 1613-0073)

The Path is the Destination – Enabling a New Search Paradigm with Linked Data

Jörg Waitelonis, Magnus Knuth, Lina Wolf, Johannes Hercher, and Harald Sack

Hasso-Plattner-Institute Potsdam,
Prof.-Dr.-Helmert-Str. 2–3, 14482 Potsdam, Germany
{joerg.waitelonis,magnus.knuth,lina.wolf,
johannes.hercher,harald.sack}@hpi.uni-potsdam.de
http://www.hpi.uni-potsdam.de/

Abstract. Today, searching the World Wide Web in most cases turns out in looking for a specific item, which means that the user should know the item in advance. In the future internet, searching for information comes closer to the notion of 'window shopping' by means of exploratory and semantic search technologies. In the course of the exploratory search process the user constantly receives new information and establishes a personalized knowledge base. To make this possible, semantic search technologies utilize domain knowledge and semantic data, as e.g., Linked Open Data (LOD), to expand and refine search results, to derive cross-references, and to reveal implicitly hidden semantic relations. Implementing semantic exploratory search requires various issues have to be solved including mapping text to semantic entities, detecting and cleaning inconsistencies in available LOD, ranking algorithms for semantic data and heuristics for recommendations, as well as appropriate visualizations of complex semantic relationships. This paper describes how LOD can be utilized to enable exploratory search systems for the future internet.

Keywords: exploratory search, linked open data

1 Introduction

When it comes to searching the World Wide Web (WWW, web), you should know what you are looking for. Today's web search engines take a query phrase consisting out of one or several key terms as input and deliver a huge set of documents that contain these key terms. The result set is presented as an ordered list in descending relevance and accuracy. To express her information needs to the search engine the user has to think of appropriate key terms. But what, if the user is not familiar with the search domain? What, if she doesn't know how to express her information needs? What, if she simply wants to know what information is available for a specific knowledge domain? These tasks are almost intractable with current web search engines. This is because of two facts: First,

the information already available in the WWW is far too large to maintain an overview of all documents concerning a certain topic. Almost nobody will look up more than the few first pages of achieved search results. Secondly, current search engines most times neglect the meaning of the document content. Thus, it is not possible to deduct nearby cross-connections towards meaningful information, which is closely connected to the user's search request and potentially the solution of his original information needs.

The emergence of semantic web technologies enables the machine understandable representation of knowledge encoded in web documents. With the Linking Open Data (LOD) initiative[1] large resources of publicly available structured data from various domains have been triplified to become interlinked RDF(S) datasets. This Linked Data provides machine understandable semantics to enable the simple deduction of cross-connections between data. Natural Language Processing (NLP) technologies, media analysis, and statistics are applied to detect semantic entities and their relationships in multimedia web documents. Taking this into account, a semantic search engine should be able not only to deliver results of higher precision and recall, but also to give suggestions on what is nearby as regards to content and meaning. Thus, truly explorative search will become possible, enabling the user to discover and to explore knowledge that is hidden in web documents, and to solve complex search tasks. The Concept of exploratory search and related work is covered with in Section 2.

But, one of the basic prerequisites for the technical realization of an efficient exploratory semantic search engine is accuracy and correctness of the underlying data. This means that if a semantic search engine is build on top of linked data, the obtained search results and exploratory recommendations can only be as good as the quality of the underlying data sources and entities that are to be connected with the web document content.

We are not the first, who have identified serious flaws in current LOD resources [3, 4], esp. in DBpedia[2] as its central hub. These flaws stem from structural, syntactic, and semantic inconsistencies, ambiguities, and missing information that have to be resolved to enable an advantage over traditional keyword based search technologies and to fully exploit the potential of exploratory semantic search. How linked data is used to enable exploratory semantic search is explained in Section 3. Flaws and deficiencies in current linked data resources esp. with regard to exploratory semantic search are discussed in Section 4. The concluding Section 5 exemplifies our efforts to cope with the problems raised and points out future work. Building on adjusted and validated linked data resources exploratory search will complement today's web search fostering the exploration of knowledge in the future internet.

[1] http://esw.w3.org/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
[2] http://dbpedia.org/

2 Exploratory Search as a New Search Paradigm

With the growing amount of information available in the WWW a few keywords typed into an input box resulting in a unidirectional list of documents have become insufficient to fulfill all of the user's information needs. The web generally supports information seeking strategies such as browsing on-the-fly, selection, and navigation by trial and error. The user's expectations from search engines have therefore turned from the pure lookup-search to more exploratory search strategies including learning and investigation [6].

Exploratory search supports users in investigating the data space in depth as well as in broadness. In keyword-based search the target is known and the process of refining the search should reach the desired target as fast as possible [13]. In contrast, exploratory search assists the user in spotting a domain along variant paths [12]. The user can move backwards and forwards on alternative search paths and can thereby access all underlying and related data.

According to Marchionini search activities can be grouped in "lookup", "learn", and "investigation" [6]. While he considers keyword based search as sufficient for a lookup search (fact retrieval, question answering etc) to learn and to investigate are exploratory search activities. Learning searches are an iterative process returning various facts and media. Investigation searches can last over an extended period of time and therefore require multiple search sessions. Furthermore, exploratory search can involve multiple people collaboratively working together as White et al. point out [13].

In the keyword based search only content, whose keywords are known can be found. If the user is not experienced in the domain of the search topic the appropriate keywords for the search are difficult to devise.

Applying facetted search the user starts with a general keyword and refines the search results in an iterative filter process. Stefaner et al. published the facetted search interface "elastic-lists" [8]. It supports searching from general to special terms, navigation by selection, and trial and error by search path backtracking. However, using facetted search only, the user mainly encounters knowledge available in the documents already achieved by the underlying keyword based search. An exploratory search additionally has to enable the user to find nearby and related topics.

Explorative search is able to discover knowledge related to the original search topic in addition to the refining process of facetted search. In contrary to the keyword-based search, exploratory search requires active user involvement in several iterations. While the result of a keyword based search is only linear, the output of an exploratory search can be multi dimensional, such as e. g. linear or clustered search results, new facets, and related topics. Therefore, new user interfaces are needed to visualize search results and data relations to assist user interaction in the exploratory search process.

3 How to Support Exploratory Search with LOD

Exploratory search comprises methods to recommend alternative search paths and to suggest of related information to the original search results. To determine these cross-connections further information that enables the exploration of the repository, semantic technologies are used to implement exploratory *semantic* search.

For exploratory semantic search, the basis for exploration among other is constituted by LOD resources and relations. To expand and refine the search results and to enable following new search paths, search queries as well as search results have to be aligned to semantic entities that are interlinked by content based relationships. This facilitates to extend the search scope by the option to investigate the semantic context, different time references, or geographical references that are related to the search query or to the original search results.

The semantic exploitation of a repository, whether it is comprising textual or multimedia data, requires the content of its documents to be mapped to corresponding semantic entities. This mapping process is denoted as *named entity recognition*. In first place, it comprises the detection of named entities in the resource metadata or in the resource itself, if represented in textual format. These named entities are extracted with the help of linguistic techniques (Natural Language Processing, NLP) and are mapped to semantic entities from LOD resources (*entity mapping*). Named entities might be mapped to various semantic entities with different meanings. These ambiguities are caused by the natural language phenomenon of polysemy and can be solved by word sense *disambiguation* based on additional contextual information [2, 5, 7, 14].

In contrast to straight RDF search engines such as 'sindice' or 'sig.ma' [10], it is now possible to search for documents *and* semantic entities at the same time. Semantic entities assigned to documents extend the capabilities of traditional keyword based search by:

- true semantic facetted browsing to filter search results,
- extension of the query string with related entities and keywords, and
- recommendations of related documents and further search suggestions by following cross connections.

Search results can be reorganized and classified by clustering their entities, or assigning corresponding documents to (super-)classes and categories of the contained entities.Compared to the rather uni-dimensional search results in keyword-based search, it is now possible to follow relations in multiple directions. For instance, Fig. 1 shows an example of three related entities in DBpedia. If the user's original search is targeted for the British author 'Aldous Huxley', the user might also be interested in the works of the authors 'H.G. Wells' or 'George Orwell', which both share the Huxley's underlying Yago [9] class and are related to him by the 'influences' relationship. Properties such as 'influences' can be used to move hand over hand through the underlying RDF graph, while exploring the repository documents.

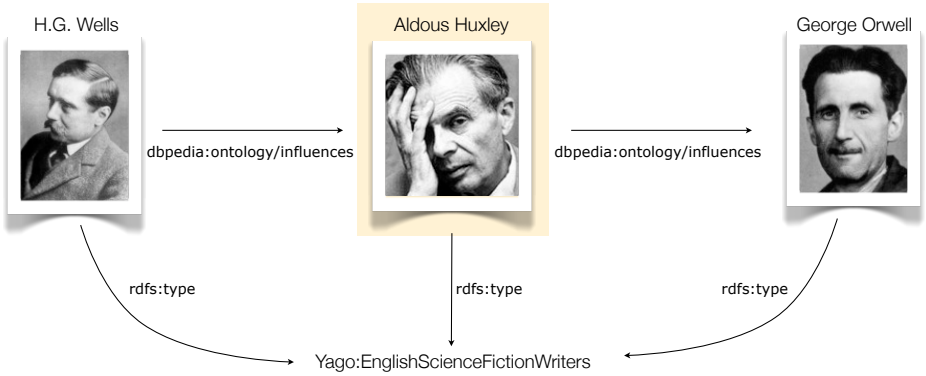

Fig. 1. Recommendations based on the DBpedia graph structure. (Images are taken from Wikipedia.)

Furthermore, entities related by time or geographical location provide a chronological or geographical classification accordingly. Time always refers to different modalities, as e. g., when a document was created, when it was published, or content based time-references. In a similar way, geographical-references can refer to different modalities, as e. g., where a document is located, where it was produced, and published, or to content based geographical references.These multimodal subtleties also require an adaption of the search interfaces. Far away from the paradigm of simple linear search result lists, new and more expressive navigational features, such as (RDF-)graph-visualizations, cluster-maps, geo-maps, and time-lines support the user in perceiving the information. Because of the high diversity w. r. t., the visualization relationships between entities, the interfaces have to be highly generic. This requires methods to prioritize and visually structure the information displayed to the user in general, according to the user's personal interests. Therefore, it is necessary to develop importance respective the relevance of related entities. For example, to visualize information about the DBpedia entity 'Albert Einstein' more than 600 facts (RDF triples) do exist. This amount of information cannot be presented to the user all at a glance. Heuristics based on statistical and semantic analysis of the underlying RDF graph structure are applied to rank related entities according their relevance [12, 11]. In addition, every user might have different preferences. Thus, the relevance rankings have to be personalized. The user's behavior can be tracked by log-file analysis. Together with the user's preferences a profile can be generated and mapped to an LOD sub graph, representing the user's interests. This enables a subjective relevance ranking and allows personalized search recommendations.

4 Issues on Using LOD to Support Exploratory Search

One essential prerequisite to achieve satisfactory search results for semantic exploratory search according to our approach as described in the previous section

is the provision of consistent and complete knowledge bases. A serious weakness of LOD data sets is their lack of data integrity and preciseness [4, 3], whereas quality strongly differs between single data sets.

In LOD there is no such thing as a consistent category system. This shortcoming makes it difficult to identify all resources of a given type. Resources often are not explicitly typed (by using `rdf:type` property) to those classes they belong to. Solving this issue becomes a forensic investigation, since in the majority of cases, types of an entity only can be deduced from the domain and range of the properties the resource is used with.

To wait for the data producers to upgrade their data is not a suitable solution. Therefore, methods are required to achieve improved semantic richness, higher quality, and completeness of linked data. Such a linked data 'washing machine' performing data cleansing on very large knowledge bases needs to implement efficient and scalable algorithms. Scalability for highly parallel and cloud computing technologies, as e. g., the HPI Future SOC Lab infrastructure[3], is mandatory to address and to solve this problem in an appropriate way.

Likewise, the performance of large-scale RDF data management has to be improved, because current triple store systems are many times less efficient compared to relational data management systems [1]. Not least for this reason, it is currently good practice to reduce processing complexity by extracting only a subset of the LOD cloud, processing it offline, and store it in a precomputed index, which in turn impairs the requirement for completeness and flexibility. One solution for this issue is the development of adaptive indexing and caching strategies.

Another issue is, how to deal with contentual gaps. The LOD cloud is far from being a comprehensive and complete mirror of available information, if ever achievable at all. However, some gaps might be closed by deductive reasoning to generate complementing RDF statements on existing data.

The success of exploratory semantic search and linked data as such heavily depends on accessible and user-friendly interfaces to visualize (intermediate) results and relations between entities. To take advantage of the rich structure of linked data, user interfaces have to display all relevant information to aggregate interrelations and to hide negligible facts depending on the context. Evaluation of such new technologies requires a reliable database of sound ground truth data, which is hard to come by. Allowing users to give immediate feedback on search results establishes a possible way to achieve this.

There is a fine line between usability and expressivity in semantic web based search. User interaction overall needs to become more intuitive, especially for building expressive queries, which have to be generated in the background by means of elementary user actions without neglecting more complex requests.

Furthermore exploratory search has to be personal. To enable personalized exploratory search results one has to keep track of provenance information beyond metadata while ensuring compliance with privacy requirements.

[3] http://www.hpi.uni-potsdam.de/forschung/future_soc_lab.html

5 Conclusion and Future Work

By harnessing the meaning of content associative, facetted, and exploratory search interfaces can be developed providing high quality search results (by means of recall and precision). The Linking Open Data (LOD) initiative has engaged various communities to share their data for sustainable usage based on semantic web technologies. However, the publicly available LOD datasets often do not meet mandatory quality requirements. The future internet will be based on semantic technologies that enable information access in a content-based way by including formal semantics in a machine understandable manner. Therefore, LOD resources require thorough analysis and subsequent bug fixing or refurbishing as we have pointed out in the previous chapters. Our current research focusses on analysis and cleansing LOD resources on structural and syntactical level as well as on semantical and contentual level. We tackle these issues with the help of the infrastructure of the HPI Future SOC Lab that provides computing resources with main memory on terabyte level to cope with the necessary large-scale processing. To deduce cross-connections to meaningful information related to the user's information needs we apply refurbished datasets of the LOD cloud to build on and propose an exploratory search paradigm. Exploratory semantic search is based on generic facets, enabling the user to better refine and broaden search queries and to provide content-based recommendations. A prototype implementation of the exploratory search featurefocussing on academic video search is publicly available[4] and subject of ongoing research. By shifting the current paradigm of web search from simple keyword based search that provides satisfactory results as long as the user knows what exactly she is looking for, to an exploratory approach, web search is becoming a quest for knowledge, guiding the user along new pathways to serendipitous findings.

References

1. Bizer, C., Schultz, A.: The Berlin SPARQL Benchmark. Int. J. Semantic Web Inf. Syst. 5(2), 1–24 (2009)
2. Bunescu, R., Pasca, M.: Using Encyclopedic Knowledge for Named Entity Disambiguation. In: Proc. of the 11th Conf. of the European Chapter of the Assoc. for Computational Linguistics (EACL2006), Trento, Italy (Apr 2006)
3. Ding, L., Shinavier, J., Shangguan, Z., McGuinness, D.: SameAs Networks and Beyond: Analyzing Deployment Status and Implications of owl:sameAs in Linked Data. In: 9th Int. Semantic Web Conference (ISWC2010). Shanghai, China (Nov 2010)
4. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the Pedantic Web. In: Proc. of the Linked Data on the Web (WWW2010) Workshop (LDOW 2010). Raleigh, North Carolina, USA (Apr 2010)
5. Li, X., Morie, P., Roth, D.: Semantic Integration in Text – From Ambiguous Names to Identifiable Entities. AI Magazine 26(1), 45–58 (2005)

[4] http://mediaglobe.yovisto.com:8080/

6. Marchionini, G.: Exploratory Search: From Finding to Understanding. Commun. ACM 49(4), 41–46 (2006)
7. Navigli, R., Velardi, P.: Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(7), 1075–1086 (Jul 2005)
8. Stefaner, M., Urban, T., Seefelder, M.: Elastic Lists for Facet Browsing and Resource Analysis in the Enterprise. In: Int. Workshop on Database and Expert Systems Applications (DEXA2008), Turin, Italy (Sep 2008)
9. Suchanek, F.M., Kasneci, G., Weikumun, G.: YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: Proc. of the 16th Int. World Wide Web Conf. (WWW2007), Banff, Canada (May 2007)
10. Tummarello, G., Delbru, R., Oren, E.: Sindice.com: Weaving the Open Linked Data. In: Proc. of the Int. Semantic Web Conference (ISWC2007). Busan, South Korea (Nov 2007)
11. Waitelonis, J., Sack, H.: Augmenting Video Search with Linked Open Data. In: Proc. of Int. Conf. on Semantic Systems 2009 (i-semantics2009). Graz, Austria (Sep 2009)
12. Waitelonis, J., Sack, H.: Towards Exploratory Video Search Using Linked Data. In: Proc. of the 2009 11th IEEE Int. Symp. on Multimedia (ISM2009). Washington, DC, USA (Dec 2009)
13. White, R.W., Roth, R.A.: Exploratory Search: Beyond the Query-Response Paradigm. No. 3 in Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers (2009)
14. Zhu, J., Zhou, X., Fung, G.P.: A Term-Based Driven Clustering Approach for Name Disambiguation. In: Proc. of the Joint Int. Conf. on Advances in Data and Web Management (APWeb/WAIM2009). Suzhou, China (Apr 2009)