Exploratory Search on the Top of DBpedia Chapters with the Discovery Hub Application

Nicolas Marie [1,2], Fabien Gandon[1], and Myriam Ribière[2]

[1] INRIA Sophia-Antipolis, Wimmics Team,
2004 Route des Lucioles, Sophia-Antipolis, 06410 BIOT
{nicolas.marie,fabien.gandon}@inria.fr
[2] Alcatel-Lucent Bell Labs,
7 Route de Villejust, 91260 NOZAY
{nicolas.marie,myriam.ribiere}@alcatel-lucent.com

Abstract. Discovery Hub is a novel application that processes DBpedia for exploratory search purpose. It implements on-the-fly semantic spreading activation and sampling over linked data sources to suggest ranked topics of interest to the user.

Motivation. Exploratory search [2] systems help users learn or investigate a topic. They provide a high level of assistance during the search task and advanced explanations about the results. In the past few years several works put in evidence the interest of using linked data, and especially DBpedia[1], for exploratory search systems. In these systems the user query is matched with the DBpedia graph. Then relevant related results are selected and ranked. These results are then presented through interfaces optimized for exploratory tasks. These systems use various methods but all depend on a partial or total results pre-processing. As it is a young research area there are many possible improvements:

- No work deals with the data freshness issue. Indeed, linked data datasets are evolving over the time. The continuous update of the data and its impact on the preprocessing need to be addressed.
- No work proposes a lightweight method that is applicable on remote SPARQL endpoints. Indeed the pre-processing phases used in the state-of-the-art works are specific to the knowledge base addressed and sometimes require a copy of the base to be performed. Consequently existing systems often only consider one data source e.g. the English-speaking version of DBpedia.

These limitations are mainly due to the preprocessing step that is strongly conditioning the type and the range of results the applications are able to retrieve. To overtake these limits, we propose a method that selects and ranks on-the-fly a meaningful subset of resources in a targeted LOD dataset. We use the term "*on-the-fly*" to stress that the method does not need any pre-processing.

[1] http://dbpedia.org

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7955, pp. 287–288, 2013.
© Springer-Verlag Berlin Heidelberg 2013

Approach. To reach our objective of an on-the-fly linked data processing we propose a novel method [3] based on a semantic spreading activation coupled with a sampling process. The spreading activation technique [1] consists in associating a value to the node(s) representing the user's interest and then spreading this value to the neighborhood iteratively with various heuristics. It was applied over RDF for various objectives (see [4]). Our approach differs as the propagation controlling pattern is a type-based semantic weight that is function of the stimulated origin node. In other words the origin node semantics plays a significant role in the distribution of activation even in "*distant*" parts of the graph:

- When a query is entered a local triple store instance is created. It imports the neighbourhood of the seed node(s) filtered with a semantic pattern taking in consideration the nodes' types. This pattern aims to concentrate the activation on a consistent subset of nodes in order to increase the relevance. The most prevalent types of the seed's neighbours are included in the pattern. For each neighbour only its deepest type(s) in the class hierarchy is/are taken in account.
- As the propagation spreads along the iterations the neighbourhoods of the most activated nodes are imported till a limit (a maximum number of triples) is reached. The semantic pattern is re-used for the imports at every iteration.
- The propagation stops when the maximum number of iterations is reached. The most activated nodes are suggested to the user in decreasing order of activation.

We performed extensive analysis on a large set of queries to understand the behaviour of the algorithm. It helped us to set its main parameters correctly in order to get a fast response time without degrading the results too much. We obtained an average response time of 2031ms with a standard deviation of 1952 ms on a set of 100.000 queries having one node as input i.e. ego-centric queries. The 100.000 nodes were selected randomly thanks to a random walker. With the proposed method it is possible to address remote data sources using their endpoints, e.g. local DBpedia chapters.

Prototype. Discovery Hub[2] is an exploratory search engine that helps its users to explore topics of interest. It uses DBpedia to perform the semantic spreading activation, render (e.g. label, description, pictures) and organize the results space (e.g. filters, facets). It also offers various explanations about the results based on DBpedia and Wikipedia. Discovery Hub is able to address localized DBpedia chapters: using the Italian data for the query "*Leonardo Da Vinci*" for instance.

References

[1] Crestani, F.: Application of Spreading Activation Techniques in Information Retrieval. Artificial Intelligence Review 11(6), 453–482 (1997)

[2] Marchionini, G.: Exploratory search: From finding to understanding. ACM (2006)

[3] Marie, N., Corby, O., Gandon, F., Ribière, M.: Composite interests' exploration thanks to on-the-fly linked data spreading activation. In: Hypertext 2013 (to appear, 2013)

[4] Rodríguez, J.M.Á., Gayo, J.E.L., Ordoñez de Pablos, P.: An Extensible Framework to Sort out Nodes in Graph-Based Structures Powered by SA Technique: The ONTOSPREAD Approach. In: International Journal of Knowledge Society Research (2012)

[2] http://semreco.inria.fr