Exploratory search on the top of DBpedia chapters with the Discovery Hub application

Nicolas Marie, Fabien Gandon, Damien Legrand, Myriam Ribière

Exploratory search on the top of DBpedia chapters with the Discovery Hub application

NicolasMarie[1,2], Fabien Gandon[1], Damien Legrand[1], Myriam Ribière[2]

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

Abstract. Discovery Hub is an exploratory search engine that helps users explore topics of interests for learning and leisure purposes.It makes use of a semantic spreading activation algorithm coupled with a sampling technique so that it does not require a preprocessing step.

Keywords: Semantic web, linked data, DBpedia, spreading activation, semantic spreading activation, exploratory search system, discovery engine

1 Linked data based exploratory search and recommendation

Exploratory search [2] systems are designed to assist users during expensive cognitive consuming search tasks such as learning or topic investigation. Theyprovide a high level of assistance during the navigation in the results space and advanced results explanations. In the past few years several works showed the interest of using linked datadatasets, and especially DBpedia[1], for resources discovery in recommender and exploratory search systems. For instance Seevl[2]is a band recommender helping the discovery of musical content and artists on Youtube thanks to a recommendation algorithm and a faceted browsing functionality. MORE[3] is a movie recommender in the form of a Facebook application that performs film recommendations thanks to DBpedia.Aemoo[4] is an exploratory search system that offers a filtered view on the DBpediagraph and gives explanations on the relations between the resources shown to the user.Yovisto[5] is a video platform offeringan exploratory search featurethat proposes a ranked list of related topic besides search results.

[1]http://dbpedia.org
[2]http://seevl.net/
[3]http://apps.facebook.com/new-more/
[4]http://wit.istc.cnr.it/aemoo
[5]http://www.yovisto.com/

All these systems match the user query with DBpediaresource(s) and perform the selection, ranking and rendering of related/similar results. They use diverse methods and depend on a partial or total preprocessing step. As it is a young research area there are many improvements possible:

- No work allows the expression of *composite* queries for exploratory search i.e. interests captured in the form of several resources ("*Claude Monet*" + "*Emile Zola*"). Using linked data to solve such queries gives the possibility to identify complex, indirect, non-trivial paths between the seed resources. Itenablesthesuggestion ofresults that are at the cross-road of different topics of interest.
- No work deals with the data freshness issue. Indeed, linked data datasetsare evolving over the time. The continuous update of the data and its impact on the preprocessing is not addressed in the state-of-the-art.
- No work proposes a lightweight method that is applicable on remote SPARQL endpoints. Indeed the pre-processing phases used in the state of the artare specific to the knowledge base addressed andoften requires a local copy of the base to be performed. Consequently the existing systems are often limited to one defined knowledge base.

These limitations are mainly due to the preprocessing step that is strongly conditioning the type and the range of results that the applications are able to retrieve. We explore the potential of an on-the-fly linked data processing for exploratory search purpose. We use the term "*on-the-fly*" to stress that the method does not need any preprocessing to produce the results. It fetches data from the SPARQL endpoint and processes it at runtime.

2On-the-fly semantic spreading activation

To reach our objective of an on-the-fly linked data processing we propose a method that is based on a semantic spreading activation coupled with a sampling phase (architecture shown on figure 1). Generally speaking, the spreading activation technique [1] consists in associating a numeric value to the node(s) representing the user's interest(s) and then spreading this value to the neighborhood iteratively with heuristics depending on the application goal. Our approach is novel as the propagation controlling pattern is a class-based semantic weight that is function of the stimulated origin node. 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 class-based semantic pattern. This pattern aims to concentrate the activation on a consistent subset of nodes in order to increase the algorithm 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 (maximum number of triples) is reached. The semantic pattern is re-used for the imports during all the process.

- 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.

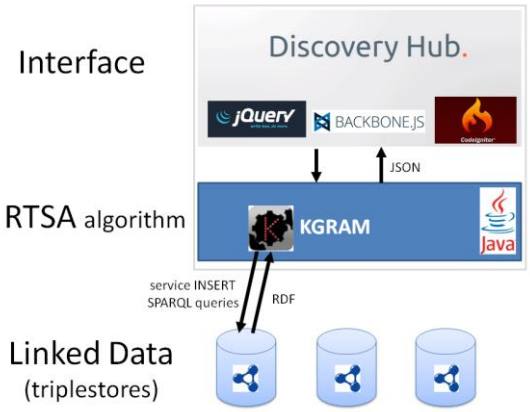


Fig.1. Discovery Hub architecture

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.During our experiment the sample had a size of 6000 triples (details in [3]).The 100.000 nodes were selected randomly thanks to a random walker. The class-based pattern and the sampling considerably lower the amount of triples needed to compute the results. Thus it is possible to run the algorithm on-the-fly, for instance on public SPARQL endpoints like the localizedDBpedia chapters (e.g. Italian, Spanish).

3The Discovery Hub web application

Discovery Hub[6] is an exploratory search engine that helps its users to explore topics of interest through an interface optimized for exploration.Discovery Hub uses the DBpedia knowledge to render the algorithm results (e.g. label, description, pictures) and organize the results space (e.g. filters, facets). See figure 2.
As the understanding of the results is very important for exploratory search systems Discovery Hub also providesseveral results explanations functionalities that help the user to understand the relation between the query-resources(s) and the results: one showing the common properties they share, one highlighting their cross-references in Wikipedia, one showing direct and indirect connections in a graph-format(see figure 3).The application proposes also many redirections to tierce platforms to extend the search process. The third-party services are proposed according to the type of the considered result e.g. music service for a *Band* or a tourism platform for a

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

Museum.Several demonstration videosare available online[7]. The results retrieved by the application were successfully evaluated thanks to a user's experimentation.

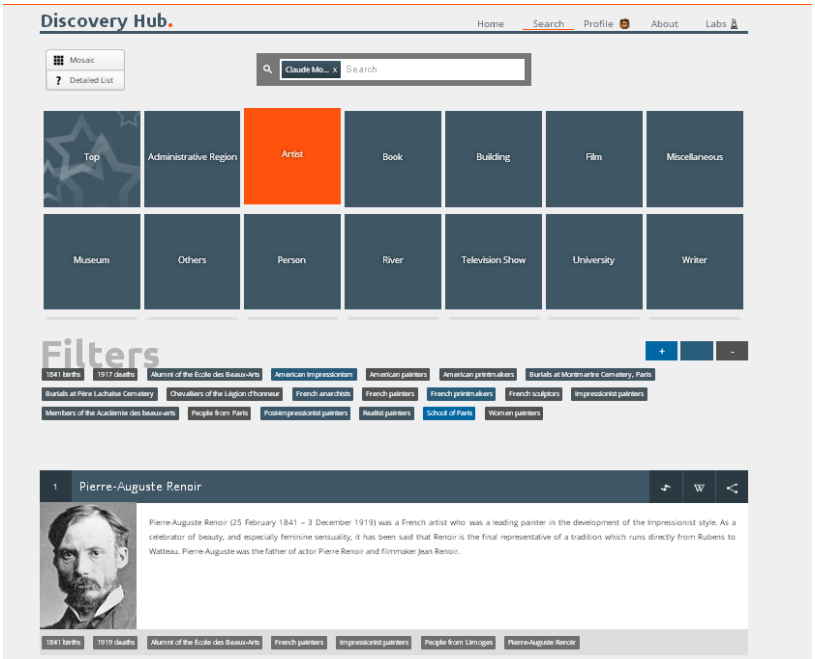
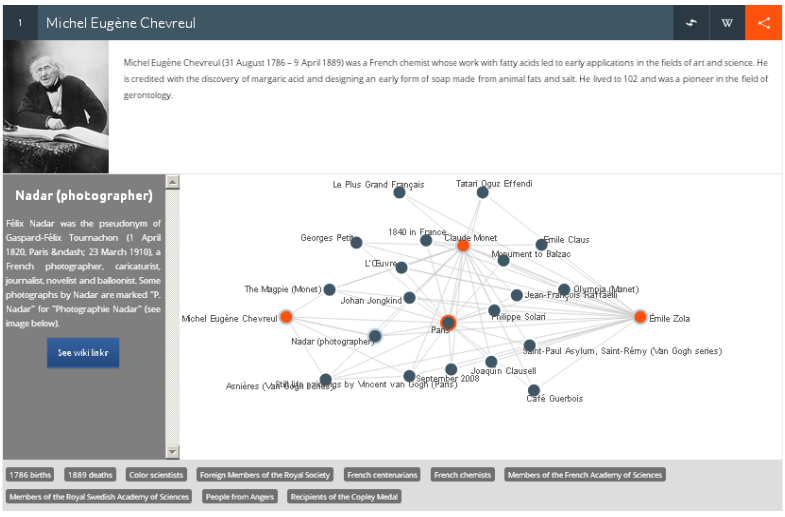


Fig.2. Discovery Hub results space for thequery *Claude Monet*.



[7]http://semreco.inria.fr/hub/videos/

Fig.3. Explanation unveiling the connections between *MichelEugèneChevreuil*(scientist) with and two seeds*Claude Monet* (painter) + *Emile Zola* (writer).

More and more local DBpedia chapters emerge[8]. Thanks to our approach it is very easy to switch from one SPARQL endpoint to another. Today Discovery Hub is able to process Czech, English, French, German, Italian, and SpanishDBpedia data.With the "*internationality*" mode the user can see the level of description of the resource in the various DBpedia chapters. This level of description corresponds to the node degree in the respective DBpedia chapters. This modehelpsthe user to select the richest data source to perform his query (e.g. use the Italian data for an Italianpainterquery). This flexibility in the choice of the SPARQL endpoint addressed is an advantage offered by the absence of preprocessing.
On demonstration we will develop an exploratory search scenario centered onthe painter *Claude Monet*. We willusethefaceted browsing featurestoexplore the results space. As these interface elements conveylotof knowledge about the results, we will show that they can be a source of inspiration during the search process,drive the user in unexpected browsing paths or help him to identify his knowledge gaps. We will give examples of the type–based redirections:Google Art Project[9] for an *Artist* resultand TripAdvisor[10] for a *Museum*one. We will show the explanations featuresfor the result*Camille Pissaro* result and how they can be combined. Then we willshowcase an example of polycentric query starting from *Claude Monet*and*Emile Zola*seeds. We want to demonstrate thatDiscovery Hub is able to identify andretrieve quickly a synthesized view on complex resources connections. The query will be composed with the "*search box*" in which the user can drag and drop resources of interest all along his navigation. Finally we will show the capacity of the application to addressremote public SPARQL endpointsby executing the*Claude Monet* query on several localized DBpedia chapters. We will finish by observing the results differences depending on theknowledge base.

4Conclusion

Discovery Hub prototype implements a semantic sensitive spreading activation algorithm coupled with a sampling technique. It allowsprocessing the linked data on-the-flyi.e. withoutany pre-processing and offers flexibility in the choice of the SPARQL endpoint and explanation mechanisms.

References

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

[8] http://dbpedia.org/internationalization
[9] http://www.googleartproject.com/
[10] http://www.tripadvisor.com/

[2] G. Marchionini. 2006. Exploratory search: From finding to understanding. *Comm. Of the ACM, 49(4), 2006.*

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