# SEMLEX - A Framework for Visually Exploring Semantic Query Log Analysis

Suvodeep Mazumdar, Khadija Elbedweihy, Amparo E. Cano, Stuart N. Wrigley and Fabio Ciravegna

OAK Group,
Dept. of Computer Science, The University of Sheffield,
Regent Court - 211 Portobello Street, S1 4DP, Sheffield, UK
{s.mazumdar, k.elbedweihy, a.cano, s.wrigley,
f.ciravegna}@dcs.shef.ac.uk

**Abstract.** With organisations, local bodies and governments now releasing large amounts of linked data, there is a great opportunity for users and software agents to look for structured information, serving their information needs. Query logs, preserving such information needs can be harvested to understand what linked data consumers are looking for. Though statistical analysis of such query logs have been employed over the years to improve performance, visualising such analyses can provide a different way of exploration that could be invaluable to researchers, developers and linked data providers for discovering hidden trends and patterns. This paper presents our approach to analyse query logs and introduces SEMLEX, a tool that facilitates visual exploration of semantic query log analysis.

**Keywords:** linked data, information visualisation, semantic query log analysis, information needs

## 1 Introduction

Over the past few decades, researchers have gained significant insights into the information needs of users of traditional search engines. However, traditional query logs limit the analysis to understanding a set of keywords and URLs, thereby often missing out on semantic context. Analysis of semantic query logs can provide much more information by analysing structure, usage and semantic context. This has become increasingly significant with the widespread adoption of linked data. Analysing query logs from publicly available data sources can provide immense possibilities as that can provide a spotlight on things that are interesting to users. Most of the previous studies have focussed on metadata statistics derived from Semantic Web search engines [1,2]. However, there has been little focus on user interfaces that enables exploration of such analyses. In this paper, we present our generic approach for:

- Analysis of semantic query logs
- Formalisation of query log analysis
- Exploration of formalised query log analysis using SEMLEX

The analysis of query logs results in the generation of RDF triples based on an ontology (QLog ontology[1]), developed for formalising semantic query logs. The RDF triples are then consumed by visualisation modules using Prefuse[2]. Users can explore such analyses using the SEMLEX (SEMantic Log EXplorer) tool, which provide users with two visualisations - Concept Graph and Predicate Transition Tree.

## 2 Query Logs Analysis

### 2.1 Modelling Query Logs

Our initial analysis of the query logs begin with identifying the key concepts that can be extracted from a query log entry. We base our model on the commonly used formatted logs called Combined Log Format[3] (CLF). Concepts such as Request IP, Request Date, Response Code, Response size and so on can be easily extracted from such log entries. These concepts are mapped to the QLog ontology. The Request string thereby extracted from a CLF log entry is then parsed to identify the exact SPARQL query. The SPARQL query is then analysed to identify different properties like types of triple patterns, joins, subjects, objects, predicates, filter types, constructs and so on. This information is then mapped to the QLog ontology. Though there is plenty of scope for improving and engineering the QLog ontology, our primary focus is to propose a methodology for analysing and consuming query logs and not a highly engineered ontology that aims to encapsulate all information extracted/analysed from query logs.

### 2.2 Consuming Query Log Analysis

SEMLEX consumes the generated RDF graph and provide users with two interactive visualisations. The visualisation modules (Concept Graph and Predicate Transition Tree) query a local triple store for all the data instances and predicates that have been identified in the query log dataset. The identified data instances are then used by the Concept Graph module to query the source endpoint for their relevant classes. The linked data endpoint is further queried with the classes to retrieve the total number of instances. This information is then aggregated to build internal data tables, which are finally rendered as nodes and edges. The nodes in the graph represent classes and the edges represent the relations among these classes. Figure 1 shows the most often queried DBpedia concepts, represented as nodes. The nodes are visually encoded using two sets of information - size (to represent how many instances are types of the class) and colour (to represent how many times the concept has been queried). By glancing through the visualisation, users can immediately identify that concepts like Person, Organisation, Work, Place contain the largest number of instances (the nodes which have larger size). Concepts like Educational Institution, University, Band, Musical Work etc are most queried for.

[1] The QLog ontology and a video of SEMLEX is available at `http://oak.dcs.shef.ac.uk/QLogAnalysis/`

[2] Prefuse Visualization toolkit, http://prefuse.org/

[3] Combined Log Format, http://httpd.apache.org/docs/1.3/logs.html#combined

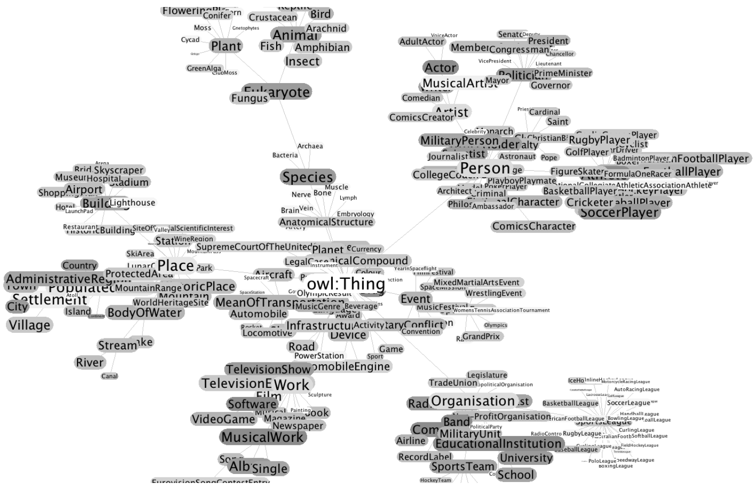

**Fig. 1.** Concept Graph - Node size represents the amount of instances (larger nodes represent more instances), color represent the amount of user interest (darker nodes represent more interest)

Another interesting feature that can be extracted by analysing query logs is how users queried for information in SPARQL queries, specially when using several predicates to connect individual triple patters. We refer to a predicate transition as a transition of a user's information need from one type of information to another. For example, if a user queries for the names and birthplaces of persons, a sample SPARQL query would be as follows:

```
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?name ?place WHERE {
     ?person dbo:birthPlace ?place.
     ?person foaf:name ?name.
}
```

The predicate transition in the query would be from dbo:birthPlace to foaf:name. If several queries reflect the same transition, then the analysis would determine this particular transition as an important one. We propose a predicate transition tree for visualising such predicate transitions. The predicate transition tree is a simple tree view, where each node is a predicate, and an edge is a link to the next used predicate. The next used predicates are ranked from top to bottom, according to their usage (most used predicate is arranged at the top). The sequence of appearance of predicates is preserved during the data extraction and analysis phase. This predicate sequence is then accumulated and a transition matrix is built which is a n X n matrix, where n is the number of predicates. Each cell (Cell A,B) signifies the transition of predicate A to B, and is represented by a number between 0 and 1. This transition matrix is read and converted to internal data tables, which are finally rendered as nodes and edges.

Figure 2 shows how predicates in DBpedia have transitioned to the next most used predicates. On providing a seed property (here, the property 'starring' has been used), SEMLEX identifies the next most used predicates by referring to the transition matrix.

This generates the next level of information (label, imdbid, stars, comment, budget...). On clicking the property ‘imdbid’ the user is presented with the next level of information (id, homepage, imdbid). This signifies that when users have used the predicate ‘imdbid’, most of the times they have used ‘id’ as the immediate next predicate, whereas ‘imdbid’ as the least used next predicate.

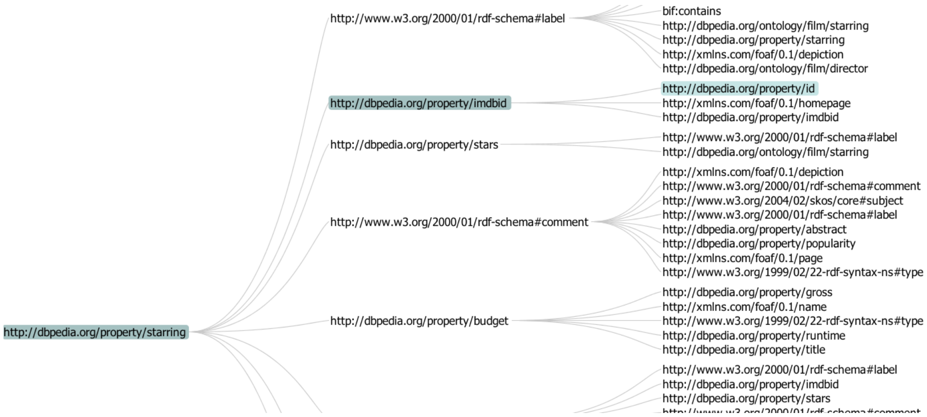
http://dbpedia.org/property/starring
http://www.w3.org/2000/01/rdf-schema#label
bif:contains
http://dbpedia.org/ontology/film/starring
http://dbpedia.org/property/starring
http://xmlns.com/foaf/0.1/depiction
http://dbpedia.org/ontology/film/director
http://dbpedia.org/property/imdbid
http://dbpedia.org/property/id
http://xmlns.com/foaf/0.1/homepage
http://dbpedia.org/property/imdbid
http://dbpedia.org/property/stars
http://www.w3.org/2000/01/rdf-schema#label
http://dbpedia.org/ontology/film/starring
http://www.w3.org/2000/01/rdf-schema#comment
http://xmlns.com/foaf/0.1/depiction
http://www.w3.org/2000/01/rdf-schema#comment
http://www.w3.org/2004/02/skos/core#subject
http://www.w3.org/2000/01/rdf-schema#label
http://dbpedia.org/property/abstract
http://dbpedia.org/property/popularity
http://xmlns.com/foaf/0.1/page
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://dbpedia.org/property/budget
http://dbpedia.org/property/gross
http://xmlns.com/foaf/0.1/name
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://dbpedia.org/property/runtime
http://dbpedia.org/property/title
http://www.w3.org/2000/01/rdf-schema#label
http://dbpedia.org/property/imdbid
http://dbpedia.org/property/stars

**Fig. 2.** Predicate Transition Tree - The figure shows that after cumulating predicate sequences of all the queries, for a particular property (e.g. imdbid), what are the other predicates (e.g. id, homepage, imdbid in descending order) used as the next predicate in one query

## 3 Demo details

The proposed demo would illustrate (1) The query log analysis process; (2) Query log analysis visualisation using SEMLEX; (3) Exploration of Semantic Web Dog Food query logs dataset to identify the information needs for SW users

**Acknowledgements** K. Elbedweihy and S.Wrigley are funded by the European Union 7th FWP ICT based e-Infrastructures Project SEALS (Semantic Evaluation at Large Scale, FP7-238975). A.E. Cano is funded by CONACyT, grant 175203. S.Mazumdar is funded by SAMULET, a project supported by Rolls Royce and Technology Strategy Board.

## References

1. M. Arias, J. D. Fernández, M. A. Martínez-Prieto, and P. de la Fuente. An empirical study of real-world sparql queries. *CoRR*, abs/1103.5043, 2011. informal publication.
2. K. Möller, M. Hausenblas, R. Cyganiak, and G. A. Grimnes. Learning from linked open data usage: Patterns and metrics. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, 2010.