Is There Anything Worth Finding on the Semantic Web?

Harry Halpin
Institute for Communicating and Collaborative Systems
University of Edinburgh
10 Crichton St.
Edinburgh, United Kingdom
H.Halpin@ed.ac.uk

ABSTRACT

There has recently been an upsurge of interest in the possibilities of combining structured data and ad-hoc information retrieval from traditional hypertext. In this experiment, we run queries extracted from a query log of a major search engine against the Semantic Web to discover if the Semantic Web has anything of interest to the average user. We show that there is indeed much information on the Semantic Web that could be relevant for many queries for people, places and even abstract concepts, although they are overwhelmingly clustered around a Semantic Web-enabled export of Wikipedia known as DBPedia.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search Process

General Terms

Experimentation, Measurement

Keywords

Search, Semantic Web, information retrieval, Linked Data

1. SAMPLING THE SEMANTIC WEB VIA QUERY LOGS

The main problem confronting of any study of the Semantic Web is one of *sampling*. As almost any large-data database can easily be exported to RDF, statistics demonstrating the actual deployment of the Semantic Web can be biased by the automated release of large, if useless, data-sets, the equivalent of 'Semantic Web' spam. Also, large specialized databases like Bio2RDF can easily dwarf the rest of the Semantic Web in size. A more appropriate strategy would be to try to answer the question: What information is available on the Semantic Web that users are actually interested in? The first large-scale analysis of the Semantic Web was done via an inspection of the index of Swoogle by Ding and Finin [4]. The primary limitation of that study was that the study was done before the release of Linked Data resources like the DBPedia, and therefore missing much Semantic Web data users might find interesting [1].

The obvious candidate for sampling what users are interested in would be look at a search engine query log. Since Semantic Web search engines are currently under development and used mostly by Semantic Web developers instead of ordinary users, the query log of a popular hypertext search engine should be used as opposed to a more specialized search engine. We use a search query log of approximately 15 million distinct queries from Microsoft Live Search. This query log contains 14,921,285 queries. Of these queries, 7,095,302 (47.55%) were unique, and corrected for capitalization, 6,623,635 (44.39%) were unique.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. ENTITIES AND CONCEPTS

The main issue confronting using a query log for discovering information on the Semantic Web is the presence of navigational and other non-informational queries. A straightforward gazetteer-based named-entity recognizer was employed to discover the names of people and places, which are called *entity queries*. The gazetteer for person names was based off a list of names maintained by the Social Security Administration and the gazetteer for place names was based on the gazetteer provided by the Alexandria Digital Library Project. A total of 509,659 queries (7.694%) were identified as either people or places by the named-entity recognizer. We employed WordNet to approximate discovering 'abstract' concepts in the query log. Queries of length one where the query had *both* a hyponym and hypernym were selected, resulting in 16,698 queries (.003%), which we call the *concept queries*. Both types of queries are plotted in logarithmic space and both appear unsurprisingly to be power law distributions in Figure 1. The α of entity queries was calculated to be 2.31, with long tail behavior starting around a popularity of 17 and a Kolmogorov-Smirnov D-statistic of .0241 [3]. The α of the queries for concepts was calculated to be 2.12, with long tail behavior starting around a popularity of 36 with a Kolmogorov-Smirnov D-statistic of .0170.

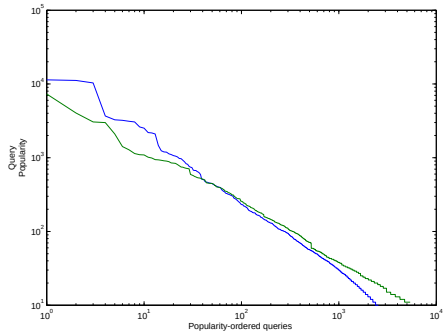

Figure 1: The rank-ordered frequency distribution of extracted entity and concept queries, with the entity queries given by green and the concept queries by blue.

3. RUNNING QUERIES AGAINST THE SEMANTIC WEB

The results of running the queries for entities and concepts against a Semantic Web search engine, FALCON-S's Object Search [2], were surprisingly fruitful. For entity queries, there was an average of 1,339 URIs (S.D. 8,000) returned for each query. On the other hand, for concept queries, there were an average of 26,294 URIs (S.D. 14,1580) returned per query. Obviously, a non-normal distribution is at hand. As shown in Figure 2, when plotted in logarithmic space, both entity queries and concept queries show a distribution that is heavily skewed towards a very large number of high-frequency results, with a steep drop-off to almost zero results instead of the characteristic long tail of a power law. For the vast majority of queries, far from having no information about content of the queries, the Semantic Web appears to have *too much data*.

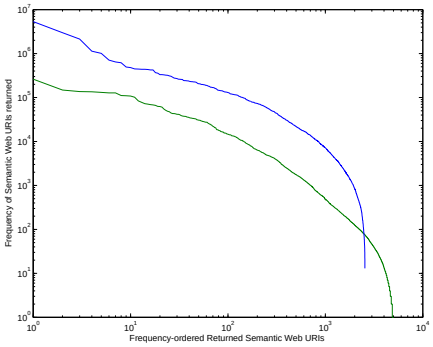

Figure 2: The rank-ordered frequency distribution of the number of URIs returned from entity and concept queries, with the entity queries given by green and the concept queries by blue.

Another question is whether or not there is any correlation between the amount of URIs returned from the Semantic Web and the popularity of the query. As shown by Figure 3, there is *no* correlation between the amount of URIs returned from the Semantic Web and the popularity of the query. For entity queries, the correlation coefficient was .0077, while for concept queries, the correlation coefficient was still insignificant, at .0125. The popularity of a query is not related to how much information the Semantic Web possesses on the information need expressed by the query: Popular queries may have little data, while infrequent queries may have a lot. This is likely due to the rapidly changing and event-dependent nature of hypertext Web queries versus the Semantic Web's preference for more permanent and less temporally-dependent data.

4. SEMANTIC WEB STATISTICS

Where is this Semantic data coming from? In order to answer this question, we restricted our analysis to the top 10 Semantic Web URI results for each query, and to distinguish this subset from all the URIs returned by the Semantic Web, we call these the *top URIs*. These top URIs totaled 70,128, which were composed of 25,400 (36.21%) concept queries and 44,728 (63.78%) entity queries. The top 10 domain names of the crawled URIs is given by Table 4. DBPedia, the export of Wikipedia to RDF, dominates the results with 83.11% of all URIs coming from DBPedia and Wikipedia [1]. The W3C is the third largest exporter of RDF with a share of 4.92%. Upon closer inspection, this is mostly due to the hosting of the

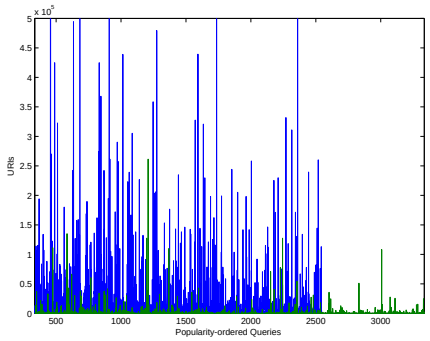

Figure 3: The rank-ordered popularity of the queries is on the x-axis, with the y axis displaying the number of Semantic Web URIs returned, with the entity queries given by green and the concept queries by blue.

54,698	78.00%	dbpedia.org
3,584	5.11%	wikipedia.org
3,448	4.92%	w3.org
1,704	2.43%	fuberlin.de
811	1.16%	cyc.com
701	1.00%	bio2rdf.org
599	0.85%	liveinternet.ru
417	0.59%	truesense.net
322	0.46%	dblp.unitrier.de
314	0.47%	ontoworld.org

Table 1: Top 10 Domain Names for Top Semantic Web URIs

lexical base Wordnet in RDF. The blog site `Liveinternet.ru` appears to export FOAF for its users.

5. CONCLUSION

In conclusion, there is a large amount of information that may be of interest to ordinary hypertext users on the Semantic Web, although there is no correlation between the popularity of queries and the availability of that information on the Semantic Web. However, the vast majority of this information is easily accessed natively in hypertext in Wikipedia or in RDF via DBPedia.

6. ACKNOWLEDGMENTS

This research and access to the query log was made possible in part by a Microsoft 'Beyond Search' award.

7. REFERENCES

[1] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, and Z. Ives. DBpedia: A nucleus for a web of open data. In *Proceedings of the International and Asian Semantic Web Conference (ISWC/ASWC2007)*, pages 718–728, Busan, Korea, 2007.

[2] Gong Cheng, Weiyi Ge, and Yuzhong Qu. Falcon-S: Searching and browsing entities on the Semantic Web. In *Proceedings of the the World Wide Web Conference*, 2008.

[3] Aaron Clauset, C. Shalizi, and M. Newman. Power-law distributions in empirical data, 2007.

[4] Li Ding and Tim Finin. Characterizing the Semantic Web on the Web. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 242–257, 2006.