INEX+DBPEDIA: A Corpus for Semantic Search Evaluation

Jose R. Perez-Aguera
Metadata Research Center.
SILS. UNC
jaguera@email.unc.edu

Javier Arroyo
GRASIA. DISIA. UCM
javier.arroyo@fdi.ucm.es

Jane Greenberg
Metadata Research Center.
SILS. UNC
janeg@email.unc.edu

Joaquin Perez-Iglesias
IR-NLP Group. UNED
joaquin.perez@lsi.uned.es

Victor Fresno
IR-NLP Group. UNED
vfresno@lsi.uned.es

ABSTRACT

This paper presents a new collection based on DBpedia and INEX for evaluating semantic search performance. The proposed corpus is used to calculate the impact of considering document's structure on the retrieval performance of the Lucene and BM25 ranking functions. Results show that BM25 outperforms Lucene in all the considered metrics and that there is room for future improvements, which may be obtained using a hybrid approach combining both semantic technology and information retrieval ranking functions.

Categories and Subject Descriptors: H.3.3[Information Systems]: Information Search and Retrieval

General Terms: Information Retrieval, Semantic Web.

Keywords: Semantic Search, Evaluation.

1. INTRODUCTION

Standard protocols for analyzing how semantic web information impacts Information Retrieval (IR) processes are severely limited. Among the best efforts to offer a standard evaluation framework for keyword-based Semantic Search is the work by [2], where TREC collections are proposed like test beds for semantic retrieval. Robust state-of-art practices focusing on keyword-based semantic search evaluations can also be found in [2].

In IR, the standard evaluation approach requires a data corpus and binary judgments in order to evaluate search engines performance. The evaluation score is automatically calculated, based on a method coming from the experiments conducted by Cyril W. Cleverdon at Cranfield University in the 60s to evaluate the performance of indexing systems [1]. The Cranfield methodology uses three basic elements: 1. documents collection, 2. a set of topics used to define queries and 3. the corresponding relevance judgments with examples of relevant and non-relevant documents for the queries generated from topics. This methodology has been successfully used for the last 30 years in the Text Retrieval Conference (TREC)[1] providing to IR community with a robust evaluation framework.

Unfortunately, the well-known body of collections for IR evaluation are not useful for Semantic Search evaluation, since they have been designed to evaluate traditional IR systems. Only, the collection from INEX[2], the XML retrieval contest, is adaptable to Semantic Search. This is due to the fact that the semi-structured formatting that characterizes XML documents is similar to the structure of RDF documents used in Semantic Search and because semantic knowledge bases like DBpedia[3] and Yago [5] have been already used in some INEX tracks.

1.1 A DBpedia + INEX evaluation framework

INEX evaluation framework provides a means for evaluating keyword-based Semantic Search systems, although some modifications must be made to adapt it to the semantic field. The main limitation is that Wikipedia, which is the source of the INEX collection, has no semantic information. To address this limitation we have mapped DBpedia to the Wikipedia version used in the INEX contest. DBpedia entries contain semantic information drawn from Wikipedia pages. In addition, DBpedia has been used in the evaluation of semantic web search engines like SEMPLORE [6], so it forms a useful corpus for this kind of tasks.

We propose to build the corpus from the intersection of DBpedia and the INEX-Wikipedia collection. The first one currently contains almost three millions of entries and the latter consists of 2,666,190 documents. The proposed corpus is made of the intersecting set of 2,233,718 documents.

With our test corpus, INEX 2009 topics and assessments are adapted to this intersection. This process has produced is 68 topics and a modified assessments file, documents that are not in the intersection have been removed. The resulting framework makes possible to evaluate the retrieval performance in a collection of RDF documents.

Our next step was to establish metrics to evaluate retrieval performance. For this purpose, we propose the use of TREC-eval software[4], which implements state-of-art IR metrics used for search engine performance evaluation [3]. These metrics are generally oriented in two main directions: the ability to retrieve relevant documents and the ability to sort them properly. We consider that these metrics are also useful to evaluate the quality of the documents retrieved in Semantic Search. In the case study shown below we have used some of them: Mean Average Precision (MAP), which is the average of the precision values measured at different recall levels; Geometric Mean Average Precision (GMAP)

[1] http://trec.nist.gov/

[2] http://www.inex.otago.ac.nz/

[3] http://DBpedia.org/About

[4] http://trec.nist.gov/trec_eval/

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

that is a variant of MAP that uses a geometric mean; Precision after X documents (P@X) which measures the precision after X documents have been retrieved; and R-Precision that measures precision after R documents have been retrieved, where R is the total number of relevant documents for a query.

2. THE CASE STUDY: LUCENE VS BM25

The evaluation framework proposed in the previous section makes possible a rigorous comparison between different ranking and indexing models. In this case, the evaluation framework is used to compare two different ranking functions: Lucene and BM25. In addition, it allows us to calculate the impact of considering documents' structure, i.e. the impact of considering the different fields of the RDF document, on the retrieval performance using LuceneF and BM25F ranking functions. We have followed SEMPLORE model to design our index [6].

The multi-field ranking approaches assign boost factors to each field in order to assign different weights depending on the field where query words occur. The boost factors used are: $text = 1.0$, $URI = 3.0$, $inlinks = 2.0$, $obj = 2.0$ and $type = 2.0$. The BM25F ranking function requires additional parameters, whose values have been $K_1 = 1.7$, $b_{URI} = 0.4$, $b_{inlinks} = 0.4$, $b_{type} = 0.4$, $b_{obj} = 0.4$ and $b_{text} = 0.3$. The fine-tuning of boost factors and the remaining parameters usually renders better performance in the retrieval task. However, it requires the use of machine learning methods and two sets of queries, one to train the machine learning method and other to evaluate it. Given that the number of queries in our evaluation framework is not high, we have not carried out the optimization and we have used values guided by our judgment. The study of the impact of fine-tuning these values remains as future work.

The set of queries in our evaluation framework is taken from the INEX 2009 contest. INEX 2009 provides judgments for 68 queries. Each query consists of three different versions of the same topic: *title*, *description* and *narrative*. For our experiments we merge the content of the three versions of each topic obtaining a more lengthy and expressive query, which fits better to Semantic Search. The resulting average length of the queries is around 30 words.

2.1 Results and discussion

Table 1 shows the results obtained by the Lucene and the BM25 ranking functions using both the plain and the multifield structure.

	MAP	P@5	P@10	GMAP	R-Prec
Lucene	.1560	.4147	.3368	.0957	.2100
LuceneF	.1200	.3971	.2971	.0578	.1632
BM25	.1746	.**.4735**	**.3868**	.1081	.2257
BM25F	**.1822**	.4647	.3824	**.1170**	**.2262**

Table 1: MAP, P@5, P@10, GMAP, R-Prec for long queries. All this measures ranges from 0 to 1

The main conclusion that can be drawn from the experiments is that BM25-based ranking functions outperform those derived from Lucene in every single measure. This is not surprising since BM25 is the state-of-the-art approach in IR. From the results it is also evident that Lucene performance dramatically decreases when structured information is used in the index. In our view, these results illustrate the problems of Lucene when dealing with structured documents that were explained in [4]. It can be seen that BM25F obtains the best performance, although the improvement over BM25 is not significant for most of the considered metrics.

A more detailed analysis of the results reveals that the distributions of the MAP values obtained by the BM25 and BM25F methods have two modes: one corresponding to the queries whose results are poor and the other corresponding to the queries whose results are quite good. This fact points to a new line of research, as it would be very interesting to identify the features of the queries that belong to each of the two underlying distributions. Confirmed evidence of the reasons behind the bimodal distribution would make possible to propose better retrieval approaches that are able to enhance the performance of the queries for which the current approaches fail to provide satisfactory results.

These conclusions can be helpful to improve the performance of Semantic Search engine implementations based on Lucene, such as Sindice, Watson, Falcons or SEMPLORE. They also highlight that there is plenty of room for collaboration between IR and Semantic Search. We firmly believe that hybrid methods combining BM25F retrieval method and semantic technology will be able to improve the obtained results.

Acknowledgments

This paper has been supported, in part, by HIVE project, funded by Institute of Museum and Library Services (IMLS). This work has been conducted in the context of the project with grant TIN2008-06464-C03-01, supported by Spanish Council for Science and Innovation. Also, we acknowledge support from the Programa de Creación y Consolidación de Grupos de Investigación UCM-BSCH GR58/08 and the very useful advices of Mounia Lalmas from University of Glasgow.

3. REFERENCES

[1] C. Cleverdon. The Cranfield tests on index language devices. pages 47–59, 1997.

[2] M. Fernandez, V. Lopez, E. Motta, M. Sabou, V. Uren, D. Vallet, and P. Castells. Using TREC for cross-comparison between classic IR and ontology-based search models at a Web scale. In *Workshop: Semantic search workshop at 18th International World Wide Web Conference*, 2009.

[3] C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA, 2008.

[4] S. E. Robertson, H. Zaragoza, and M. J. Taylor. Simple BM25 extension to multiple weighted fields. In *CIKM '04*, pages 42–49, 2004.

[5] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A large ontology from Wikipedia and WordNet. *Web Semantics*, 6(3):203–217, 2008.

[6] H. Wang, Q. Liu, T. Penin, L. Fu, L. Zhang, T. Tran, Y. Yu, and Y. Pan. Semplore: A scalable IR approach to search the web of data. *Web Semantics*, 7(3):177–188, 2009.