Integrated Retrieval over Structured and Unstructured Data

Qiuyue Wang[1,2], Jinglin Kang[1]

[1] School of Information, Renmin University of China,
[2] Key Lab of Data Engineering and Knowledge Engineering, MOE,
Beijing 100872, P. R. China
qiuyuew@ruc.edu.cn, 402817493@qq.com

Abstract. We report our experiment results on the INEX 2012 Linked Data Track. We participated in the ad hoc and jeopardy tasks. As the new data collection on INEX 2012 Linked Data Track features a combination of unstructured and structured data, our first attempt is to investigate different strategies of combining the retrievals over structured and unstructured data, and compare the combined approaches with the traditional unstructured ones. In this paper, we discussed three types of combination strategies and we experimented two of them on the track. The experiment results show that ……

1 Introduction

Though Web is best known as an enormous collection of unstructured documents, it also contains a huge amount of structured data, like HTML tables, data stored in Deep Web databases, increasingly published RDF data due to the efforts of Linked Data community, and so on. With more and more structured data became accessible to end users, more intelligent search on the Web is expected. There are increasing interests on semantic search on the Web, i.e. leveraging the semantics in structured data to improve the Web search.

The new data collection of INEX 2012 Linked Data Track is a fusion of Wikipedia articles and their corresponding RDF data from DBpedia and YAGO2. Each Wikipedia article corresponds to an entity/resource in DBpedia and YAGO2, while DBpedia and YAGO2 contain structured data extracted from each article, e.g. properties of entities and relationships with other entities. It can be viewed as an integrated collection of unstructured and structured data covering a wide range of topics. With such a data collection, we intended to investigate different strategies of combining retrievals over unstructured and structured data so that the performance would be better than that of unstructured retrieval or structured retrieval only. Basically, there are three types of combination strategies. (1) **Parallel combination**. Retrieve the structured and unstructured data separately, and then combine the two result lists. (2) **Unstructured-structured serial combination**. Retrieve the unstructured data first. The top-k results, which correspond to entity nodes in the RDF graph, then spread their activations over the RDF graph. Thus, some relevant results which do not contain query terms may be retrieved. (3) **Structured-unstructured**

serial combination. Retrieve the structured data first. The top-k returned entities or subgraphs are then analyzed so that the original query could be better understood. For example, the query is expanded with more effective terms, or is reformulated in terms of related entities and so on. The newly transformed query is then used to retrieve the unstructured data more accurately.

Due to the limit of time, we only experimented on the first two strategies. Firstly, we indexed the unstructured and structured data separately, and used language modeling approaches to retrieve them respectively. We treat the unstructured run as our baseline. Then we combined the unstructured and structured runs using weighted sum approach. Secondly, we used the unstructured run as the input to the algorithm of spreading actions on the RDF graph, and submitted the new ranked list of results after spreading activation.

The results show that ……

2 Combined Retrieval Strategies

In this section, we first present the retrieval models that we used to retrieve unstructured data and structured data respectively, and then discuss different strategies of combining the retrievals over unstructured and structured data.

Given a keyword query and unstructured document collection, we can employ any traditional IR models to retrieve relevant documents. In this paper, we use the language modeling approach since it has the state of art performance among other retrieval models.

For a structured data collection, there are various approaches proposed to look for relevant answers for a keyword query [1]. One of the key problems in structured retrieval is that return units are not predefined as in document retrieval. There are various ways to generate all possible results. Then the results are ranked using either traditional content-based models, e.g. TF-IDF, vector space model, or content-structure-based ranking models. However, there is still lack of a general evaluation campaign for comparing all these retrieval models for keyword search on structured data. So it is very hard to draw any conclusions on these various approaches. In this paper, we simply define the retrieval units of structured retrieval to be entities, which actually correspond to Wikipedia articles in the collection of Linked Data Track. To retrieve entities on RDF graphs, we first aggregate all information about an entity together, i.e. the entity's properties, subjects, objects, etc., and index it as a pseudo document identified by the entity's ID. Then we employ the language modeling approach to rank each pseudo document with respect to the given keyword query.

2.1 Parallel Combination

2.2 Unstructured-Structured Serial Combination

2.3 Structured-Unstructured Serial Combination

3 Experimental Results

Due to the limit of time, we only experimented on the first two strategies. In this section, we discuss the experiment results on the INEX 2012 Linked Data Track.

3.1 Implementation

We indexed the unstructured and structured data separately both using Indri with Krovetz stemmer and a short stop word list *{a, about, an, and, as, at, by, in, of, on, or, that, the, to}*. Remember that we generate a pseudo document for each entity in the structured data set, and index this pseudo document for the entity.

3.2 Results

The evaluation results have not been released by the time when the author wrote this abstract.

4 Conclusions and Future Work

5 Acknowledgements

The research work is supported by the “HGJ” National Science and Technology Major Project of China under Grant No. 2010ZX01042-001-002-002.

References

1. Y. Chen, W. Wang, Z. Liu, and X. Lin, Keyword search on structured and semi-structured data. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (SIGMOD '09), Carsten Binnig and Benoit Dageville (Eds.). ACM, New York, NY, USA, 1005-1010.
2. C. Rocha, D. Schwabe, and M. P. Aragao, A hybrid approach for searching in the semantic web. In *Proceedings of the 13th international conference on World Wide Web* (WWW '04). ACM, New York, NY, USA, 374-383.
3. M. Fernandez, V. Lopez, M. Sabou, V. Uren, D. Vallet, E. Motta, and P. Castells, Semantic Search Meets the Web. In *Proceedings of the 2008 IEEE International Conference on Semantic Computing* (ICSC '08). IEEE Computer Society, Washington, DC, USA, 253-260.