Hybrid Search Ranking for Structured and Unstructured Data*

Daniel M. Herzig

Institute AIFB, Karlsruhe Institute of Technology, Germany
herzig@kit.edu

Abstract. A growing amount of structured data is published on the Web and complements the textual content. Searching the textual content is performed primarily by the means of keyword queries and Information Retrieval methods. Structured data allow database-like queries for retrieval. Since structured and unstructured data occur often as a combination of both, are embedded in each other, or are complementary, the question of how search can take advantage of this hybrid data setting arises. Of particular interest is the question of how ranking as the algorithmic decision of what information is relevant for a given query can take structured and unstructured data into account by also allowing hybrid queries consisting of structured elements combined with keywords. I propose to investigate this question in the course of my PhD thesis.

1 Introduction

Currently, an increasing amount of structured data is published on the Web according to the Linked Data principles. This structured data supplements the textual, unstructured data already available on the Web and thereby provides the basis for new ways of searching the Web. The structured data is available in several ways. On the one hand there are data sets available as purely structured RDF independent from a text base, and on the other hand there is structured data embedded directly in textual data via RDFa or data extracted from texts. Taking advantage of this heterogenous environment promises to improve search by making it possible to answer more kinds of information needs, because some information needs benefit greatly from structured data, e.g. "What is the population of Berlin?". Here, the answer is a fact assertion, whereas other information needs are better addressed with textual documents, e.g. "Why did Heinrich von Kleist commit suicide?", where a potential answers might be his suicide note, if at all, but certainly not a fact assertion. Moreover, texts can hold sentiments, preferences and opinions, which are often supported by facts and data. Therefore, a hybrid data scenario holds also the possibility to examine the retrieval of opinions or different views on a topic and the facts supporting them. Thus far, document and fact retrieval are often regarded as two separate disciplines and

* This work was supported by the German Federal Ministry of Education and Research (BMBF) under the iGreen project (grant 01IA08005K).

G. Antoniou et al. (Eds.): ESWC 2011, Part II, LNCS 6644, pp. 518–522, 2011.

the combination of both for search is not yet investigated in a satisfying way[1]. This thesis is situated between these two disciplines and combines them on the data and on the query level. We call this scenario *Hybrid Search*. However, search comprises the entire technical spectrum from indexing to the user interface. This thesis concentrates on ranking, which is a core method of search and crucial for its effectiveness. The goal of this thesis is to investigate a unified ranking framework for hybrid search as the search on structured and unstructured data with queries consisting of keywords and structured elements. The question this thesis addresses is how structured and unstructured data can be used to improve search and how hybrid queries can be answered on hybrid data.

2 Problem Definition

This thesis addresses the problem of ranking on hybrid data for hybrid queries. The frame of this research problem is defined by the following data and query model: The proposed data model follows the RDF data model with Named Graphs[1] and is represented as a graph $G(R, L, E_R, E_L, E_G, \hat{G})$ consisting of resource nodes R, edges E_R connecting resource nodes, edges E_L connecting resource nodes to literal nodes L, and edges E_G connecting resource nodes to Named Graphs $\hat{G}$, which are graphs $\hat{G}(R', L', E'_R, E'_L)$ consisting of subsets of the elements of G, e.g. $R' \subset R$. Textual data is integrated following the same modeling paradigm and using the already mentioned modeling constructs. Each text entity is represented by a resource of the type *textual document*. This resource has one edge, labelled *content*, pointing to a literal node holding the textual information. In a later stage, there can be more edges providing more fine grated distinctions, such as *headline*, *paragraph*, etc. All triples comprised by the textual information of one textual entity form a Named Graph $\hat{g} \in \hat{G}$, as illustrated in Figure 1 by the dashed circle. The data model is a simplified RDF model with Named Graphs and allows to use RDF data easily.

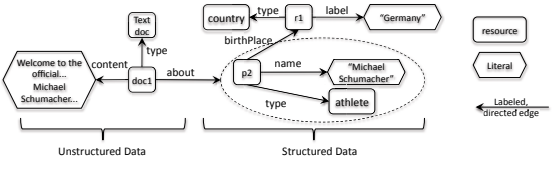

Fig. 1. Illustration of the data model. A textual document on the left side and structured data on the right side.

Queries to this data model should have a seamless flexibility ranging from purely textual keyword queries, over hybrid queries, to crisp structured queries. A hybrid query q can consist of a *structured* part q_s and a *textual* part q_t, i.e. $q = q_s \wedge q_t$. If one part is empty, the query is either purely textual or purely

[1] Named Graphs: `http://www.w3.org/2004/03/trix/`

structured. The structured part q_s follows the SPARQL query language and is a set of graph patterns, $q_s = \{q_{s1}, q_{s2}, ...\}$. The textual part q_t allows to associate a keyword query kw to each variable, i.e. $q_t = \{q_{t_i} | q_{t_i} = (x_i, kw), x_i \in Var(q)\}$. For example, assume the information need: *"Formula One drivers who moved to Switzerland"*[2], which is illustrated in Figure 2. The result to such a query are bindings to the distinguished variables. This model allows to represent purely structured, hybrid, and purely textual queries. A purely textual query, i.e. simple keyword query, would be the query in Fig. 2 without line (1). This query model is a close adaptation of the model by [3].

```
Select ?x where {
   ?x   rdf:type ns:FormulaOneRacer          # (1)
   ?x   {moved to Switzerland} }      # (2)
```

Fig. 2. Illustration of the query model, consisting of the a structured part, i.e. triple patterns (1) and unstructured part, i.e. keyword patterns (2).

3 State of the Art

Related fields of research are IR, in particular Web IR, ranking of structured data and databases, and the already existing work on hybrid. This section briefly outlines related work of these fields to the proposed approach in the Section 4.

Ranking originated in the IR community, where the retrieval units are textual documents. The main notions for ranking are descriptive statistics about the documents and the entire document corpus, such as term frequency. One of the most used algorithm in this line is BM25[4]. Language model approaches are increasingly applied, because of their formal basis in probability theory. Therefore, language models will be the basis for the proposed ranking approach. The work by [5] are of particular interest, since it builds on language models for structured retrieval, combines keywords and structured queries and addresses structural mismatches between query and data structure. However, structure means here the document and sentence structure and not a graph based data model. Ranking for Web search deals not just with fixed document corpora, but with the entire scale of the Web. However, it can take advantage of the link structure of the Web. Exemplars using this link analysis are foremost the well known works by [6] and [7]. Translating the idea of [6,7] for ranking data on the Web has been studied by [8,9]. Also concepts of XML retrieval [10] as the retrieval of semi-structured documents needs to be addressed. The parallels are here that elements in the XML scenario are similar to resources in ours. Ranking for databases draws on the advantage that the data schema is strict and rigid, which is not the case in our setting. Still, the idea of correlations[11] between values needs to be investigated. Combing text and structured data, i.e. hybrid approaches, such as [12], which uses a domain ontology for retrieving documents, or [13], which retrieves documents as well as factual assertions. However, the data setting is different to ours. Most notably is the approach by [3], which could be

[2] Topic GC-2009-49 taken from [2].

used as a reference in an evaluation, because it supports keyword augmented structured queries similar to ours. However, the approach does not take documents into account, and is centered around triples as the atomic units, where as our proposed approach regards entities, i.e. URIs of *subjects* respectively *objects*, as the atomic pieces.

4 Proposed Approach

Starting point are retrieval methods similar to [3,5] applied to a hybrid scenario, because they have proven to be applicable for similar settings and are the state of the art in IR. Following the idea of language models, we rank result graphs according to the probability of being an result graph g to the given query q, i.e. $P(g|q)$. The structured part of the query is regarded as a constraint for the beginning and can be relaxed later. It fulfills the purpose of selecting candidate results. Since q_s determines the shape of the result graphs, all possible graphs share the same structure. Therefore, the rank of a result depends only on the aspects, which differentiate the results, i.e. the bindings to the variables and their relations to q_t. Therefore, we can reduce the ranking to $P(g|q) \propto \prod_{i=1}^{n} P(q_i|x_i)$, with $q_i = q_{t_j} \wedge q_{s_k}, x_i \in q_{t_j}, g_{s_k}$, the keyword and triple patterns associated to variables x_i.

$$P(g|q) \propto P(g) \cdot P(q|g) \propto P(g) \cdot \prod_{i=1}^{n} P(q_i|x_i) \tag{1}$$

$P(q_i|x_i)$ is computed in two ways depending whether q is a purely structured query or not. If it is purely structured, the query is crisp and we assume that the information need is entirely captured. The ranking is then based on the popularity of the resulting URIs. If a textual part is present, the information need is rather imprecisely captured making it necessary to rank the results by measuring the relation of each keyword k of the textual part to the corresponding variable, see equation (2).

$$P(q_i|x_i) = \begin{cases} \prod_{k \in q_{t_i}} \alpha P_t(k|x_i) + (1-\alpha) P_t(k) & \text{if } q_t \neq \emptyset \\ \prod_{x_i} P_s(x_i) & \text{if } q_t = \emptyset \end{cases} \tag{2}$$

If a textual part is present, several models for computing $P_t(k|x_i)$ will be investigated: Starting with a simple one, which takes all textual information of x_i as one bag-of-words and a more fine grained one, which takes the edges from x_i to neighboring nodes into account.This ranking model is an initial model for the study of search in the hybrid scenario. Possible future directions of research are to extend this model and integrate more of the semantics provided by the underlying data.

5 Evaluation Methodology

The widest acceptance in IR for evaluating ranking approaches has the so-called *Cranfield methodology*[14]. It provides well studied grounds and will be the basis of the evaluation in line with [15]. However, the setting needs to be adapted to the hybrid scenario. This can be done by adding structured elements to the keyword

queries of [15] and by using datasets, which are a combination of structured and unstructured data, e.g. the combination of *Wikipedia* and *dbpedia*.

6 Conclusion

The goal of this thesis is to investigate a unified ranking methodology for search on hybrid data using adaptive queries. The proposed approach builds on a graph based data model, which is compatible to RDF and incorporates textual documents. The query model allows seamless querying ranging from purely textual queries, to hybrid queries, and to purely structured queries. The ranking approach builds methodologically on language models. The evaluation methodology uses existing standards from the IR community, if applicable, but needs to be adapted to the hybrid context. The question this thesis addresses is how the combination of structured and unstructured data can be used to improve search.

References

1. Weikum, G.: DB & IR: both sides now. In: SIGMOD (2007)
2. Santos, D., Cabral, L.M.: GikiCLEF: crosscultural issues in an international setting: asking non-English-centered questions to wikipedia. In: CLEF (2009)
3. Elbassuoni, S., Ramanath, M., Schenkel, R., Sydow, M., Weikum, G.: Language-model-based ranking for queries on RDF-graphs. In: CIKM 2009 (2009)
4. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval 3(4), 333–389 (2010)
5. Zhao, L., Callan, J.: Effective and efficient structured retrieval. In: CIKM (2009)
6. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. In: ACM-SIAM, San Francisco, California, United States, pp. 668–677 (1998)
7. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: WWW, Brisbane, Australia, pp. 107–117 (1998)
8. Harth, A., Kinsella, S., Decker, S.: Using naming authority to rank data and ontologies for web search. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 277–292. Springer, Heidelberg (2009)
9. Delbru, R., Toupikov, N., Catasta, M., Tummarello, G., Decker, S.: Hierarchical link analysis for ranking web data. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010. LNCS, vol. 6089, pp. 225–239. Springer, Heidelberg (2010)
10. Lalmas, M.: XML Retrieval. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, San Francisco (2009)
11. Chaudhuri, S., Das, G., Hristidis, V., Weikum, G.: Probabilistic ranking of database query results. In: VLDB, pp. 888–899 (2004)
12. Rocha, C., Schwabe, D., Aragao, M.P.: A hybrid approach for searching in the semantic web. In: World Wide Web, WWW 2004, New York, NY, USA (2004)
13. Bhagdev, R., Chapman, S., Ciravegna, F., Lanfranchi, V., Petrelli, D.: Hybrid search: Effectively combining keywords and semantic searches. In: Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.) ESWC 2008. LNCS, vol. 5021, pp. 554–568. Springer, Heidelberg (2008)
14. Cleverdon, C.: The CRANFIELD Tests on Index Langauge Devices. Aslib (1967)
15. Halpin, H., Herzig, D.M., Mika, P., Blanco, R., Pound, J., Thompson, H.S., Tran, D.T.: Evaluating ad-hoc object retrieval. In: IWEST 2010, ISWC (2010)