Kuphi – An Investigation Tool for Searching for and via Semantic Relations

Michael Färber*, Lei Zhang**, Achim Rettinger

Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
{michael.faerber,l.zhang,rettinger}@kit.edu

Abstract. In this work, we present a new process-oriented approach for information retrieval called *Kuphi*. It is intended for investigating entities and their semantic relations to other entities in text documents. We extend the traditional search capabilities which are based on the bag-of-words model in the following way: Starting with a keyword search for a specific entity, the user can not only search for appearances of this entity in the text documents; she can also search via user-specified relations of the queried entity to other entities for these associated entitites in the text. The user has the possibility to search indirectly for manifestations of these relations. Due to cross-lingual annotation, we allow the query language to be different from the language of the documents. We demonstrate our approach with DBpedia as knowledge base and news texts gathered from RSS feeds.

Keywords: Semantic Search, Document Ranking, Cross-lingual Annotation.

1 Introduction

The idea of annotation-based document retrieval is that queries and documents with additional annotations enhance document search. A study [1] showed that the quality of information extraction and, therefore, annotation, has a high impact on the semantic search performance.

Kandogan et al. [2] present an approach called *Avatar Semantic Search* where keyword queries are transformed into one or more queries over the used structured data set. They limit themselves to a specific domain with a few concepts such as email. In [3] an ontology-based scheme for a semi-automatic annotation of documents and a retrieval system is presented. The ranking is based on an adaptation of the traditional vector space model taking into account weights for annotations.

* This work is supported by the German Federal Ministry of Education and Research (BMBF) under grant 02PJ1002 (SyncTech).

** The authors acknowledge the support of the European Community's Seventh Framework Programme FP7-ICT-2011-7 (XLike, Grant 288342).

The work we present here can be dedicated to research in this area. It provides a significant new search paradigm. It is intended for investigating entitites and their relations written in documents. On the one hand it is based on the observation that keyword search has proven to be the most intuitive way for end users to satisfy their information needs. On the other hand it is based on the fact that nowadays there are many news portals which provide a vast amount of textual information which needs to become accessible by end users via *targeted search*. *Targeted search* means the user can (i) search for specific entitites occurring in text documents; (ii) she can search for relations of these entities to other entities (expressed in prose text); and (iii) she can search via these specified relations for other entities which stand in relation to the search entity in a specific way.

From a technical point of view, we make the following contributions: (1) *Kuphi*[1,2] exploits the knowledge base (KB) semantics during the document retrieval process, which includes the steps *text annotation*, *keyword matching*, *query refinement* and *document ranking*. The rich semantics of DBpedia as used KB are firstly used to obtain a semantic representation of the documents. During the online search process, the KB is also used to infer the semantics of queries. Based on the various semantic interpretations that can be found for the ambiguous query, users can choose the refinement that match the intent. (2) The main difference to existing works is our strong emphasize on semantic relations during the ranking of documents. One the one hand, we use them to capture the semantic focus of documents and to rank them according to how well the query matches the focus. Also, they are used for manual weighting, a mechanism we introduce for the users to influence the ranking during the search process. (3) Our semantic search system is designed for cross-lingual search. The user can select the query language and also the language of the documents to search for. This is enabled by using the huge cross-lingual lexica called xLiD [4]. This feature is especially interesting in case documents about a topic are only available in other languages than the query language. For instance, breaking news are at the beginning often written in the local language.

2 Document Search with *Kuphi*

We present a process-oriented approach to document search, which can start with a rather vague information need that becomes more concrete during the process. It assists users in specifying and addressing their information needs through several steps of a search process, as shown in Fig. 1.

In the following, we first discuss the preprocessing step, namely *text annotation*, before we focus on the on-line steps where the user is involved in.

Text Annotation. This preprocessing step is performed to enrich documents with contexts that are linked to KBs to help to bridge the ambiguity of natural language text and precise formal semantics captured by KBs. For our

[1] *Kuphi* means "Where?" in Zulu language.
[2] The demo is available at `http://km.aifb.kit.edu/services/kuphi/`.

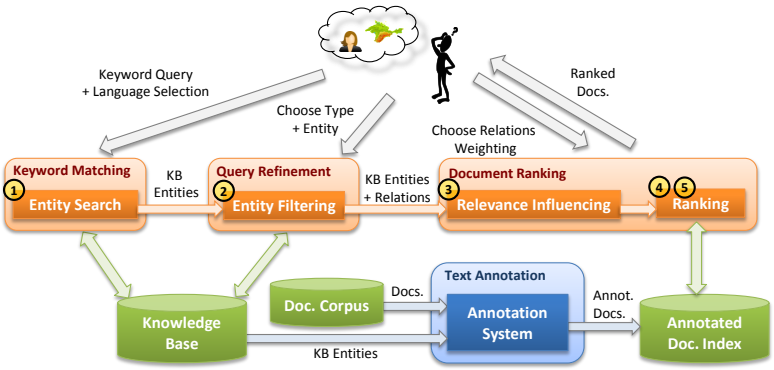

Fig. 1: The document search process with *Kuphi*. The numbers represent the different UI components as shown in Fig. 2.

demonstration, we use DBpedia as formal KB containing entities and semantic relations between them.

For linking textual mentions – also called *surface forms* – to KB entities, we apply a huge lexica data set called xLiD [4]. In this data set, for all DBpedia entites we provide a list of potential textual mentions. For the used data set we retrieved this mention list by extracting anchor texts from an English Wikipedia dump of July 2013.

An example of an annotated sentence in a document is the following:

```
[[John McCain|Senator John McCain]] was one of several senior
[[Republican Party (United States)|Republicans]] opposed to the
[[International Monetary Fund|IMF]] measures.
```

Here wiki-syntax was used to separate each surface form from its actual KB representation, the entity. For instance, the mention "IMF" was correctly disambiguated and linked to the entity "International Monetary Fund". Note that in other document languages than English we also use our xLiD data set, but link directly to the English DBpedia entities. The canonicalized DBpedia, hence, is used as hub for all provided languages.

Given the documents and their linked entities stored in the index, we now discuss the online steps involved in our search process. We will use the query "Krim" (German for the Crimea) as an example throughout the search process. Our user wants to investigate about the current Crimea crises and is especially interested in documents in English which adress the capital and the ethnical groups of the Crimea.

Keyword Matching. While keyword queries are simple and intuitive, they can be highly ambiguous. Even the phrase "Krim" could stand for the Crimea, the peninsula, or other entities such as Mathilde Krim, a medical researcher. Using the bag-of-words model that relies on term information only, semantic ambiguities of these kind are difficult to address. Thus, even when the precise query intent is given, e.g., "Krim", it is difficult to tell which of the retrieved documents are actually relevant.

Our online search process starts with a possibly ambiguous keyword query (see Frame 1 in Fig. 2). The ambiguity is resolved with the help of the user.

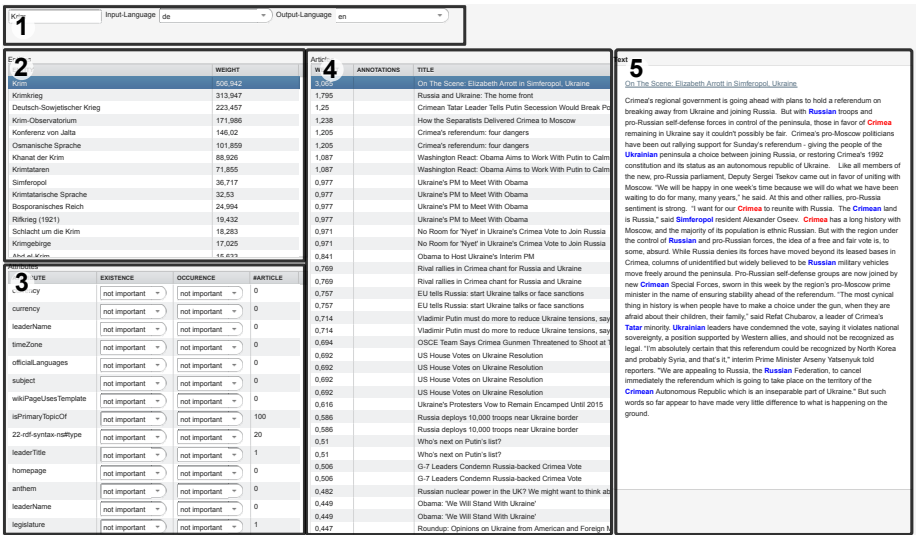

Fig. 2: Screenshot of our semantic search system *Kuphi* with frames indicating the different steps in the search process.

Instead of retrieving documents, our approach first finds entities from the KB with labels matching the query keywords. These entities represent different semantic interpretations of the query and are, thus, employed in the following steps to help the user to refine the search and influence document ranking according to the intent. The ranking of the entities is performed by means of pre-computed DBpedia PageRank values and the prior probabilities of the surface forms.

Query Refinement. By displaying all possible interpretations of the query (cf. Frame 2 in Fig. 2), the user selects the intended entity she wants to search for in the documents.

Document Ranking. Afterwards, the documents containing the queried entity are retrieved from the index[3] (cf. Frame 4 and 5 in Fig. 2). We observe that while annotated documents have different links to the KB, they generally share the following structure pattern: Every document is linked to a set of entities. A subset of these entities are connected via relations in the KB, forming a graph. Based on these entities and relations, a document can also be conceived as a graph containing several connected components. The largest connected component represents the main focus of the document. For instance, the example document (written in English) contains the entity *Crimea* and other entities such as *Crimean Tatars*, *Ukrainians*, *Russians*, and *Simferopol*. From the KB we can infer the relations between these entities. In our concrete example we have the relations *ethnic group* and *capital* of the entity *Crimea*.

Focus-Based Ranking: Leveraging this structure pattern, we incorporate the following intuition into our ranking scheme: Given two documents d_1 and d_2 retrieved for the entity / query intent e, d_1 is more relevant than d_2 if it focuses more on e than d_2 does, i.e., when the largest connected component of d_1

[3] We use Apache Lucene to index the documents together with their annotations.

containing e is larger than the largest connected component of d_2 containing e. For example, the document containing the connected entities *Crimea*, *Simferopol* and *Tatars* is more relevant than the document containing *Crimea* only.

Relation-Based Weighting: We enable the user to influence the document ranking by adjusting the weights of entity relations to obtain a personalized document ranking (see Frame 3 in Fig. 2). For this, the chosen entity is shown and extended with relations to other entities retrieved from the KB. For instance, if a user would like to obtain information about the *Crimea* and its capital, she would increase the weight of the *capital* relation and give no weight to other relations, so that the latter are not considered for ranking. This relation-based search capability is especially useful when the user does not know the capital by its name. Furthermore, the user can weight both the existence of a relation and the number of its occurrences in the document[4]. This differentiation separates the one scenario where the user is interested in obtaining more detailed information about the relation itself from the other, where users are interested in the quantity of relations. In this way, varying intents, such as "the number of different ethnical groups on the Crimea" and "one ethnical group of the Crimea", can be distinguished.

3 Conclusions

We presented a process-oriented approach called *Kuphi* for searching for entities and relations in documents. We discussed that the semantics captured by the KBs, especially semantic relations, can be exploited in this process to allow the information needs of the user to be specified and addressed on the semantic level. Based on DBpedia as KB and on news texts gathered from RSS feeds in different languages we demonstrated the practical benefit. In the future, we will advance the query capability of *Kuphi* to support information needs involving several entities.

References

1. Chu-Carroll, J., Prager, J.: An experimental study of the impact of information extraction accuracy on semantic search performance. In: Proceedings of the CIKM '07, New York, NY, USA, ACM (2007) 505–514
2. Kandogan, E., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., Zhu, H.: Avatar semantic search: a database approach to information retrieval. In: Proceedings of SIGMOD '06, New York, NY, USA, ACM (2006) 790–792
3. Castells, P., Fernandez, M., Vallet, D.: An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. IEEE Trans. on Knowl. and Data Eng. **19**(2) (2007) 261–272
4. Zhang, L., Färber, M., Rettinger, A.: xLiD-Lexica: Cross-lingual Linked Data Lexica. In: Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014), to appear (May 2014)

[4] Technically, not the number of the relations, but the number of the objects of the relations is weighted.