Exploiting Semantic Annotations for Entity-based Information Retrieval

Lei Zhang[1], Michael Färber[1], Thanh Tran[2], and Achim Rettinger[1]

[1] Institute AIFB, Karlsruhe Institute of Technology, Germany
[2] San Jose State University, USA
{l.zhang,michael.faerber,rettinger}@kit.edu,
{ducthanh.tran}@sjsu.edu

Abstract. In this paper, we propose a new approach to entity-based information retrieval by exploiting semantic annotations of documents. With the increased availability of structured knowledge bases and semantic annotation techniques, we can capture documents and queries at their semantic level to avoid the high semantic ambiguity of terms and to bridge the language barrier between queries and documents. Based on various semantic interpretations, users can refine the queries to match their intents. By exploiting the semantics of entities and their relations in knowledge bases, we propose a novel ranking scheme to address the information needs of users.

1 Introduction

The ever-increasing amount of semantic data on the Web pose new challenges but at the same time open up new opportunities for information access. With the advancement of semantic annotation technologies, the semantic data can be employed to significantly enhance information access by increasing the depth of analysis of current systems, while traditional document search excels at the shallow information needs expressed by keyword queries and the meaningful semantic annotations contribute very little. There is an impending need to exploit the currently emerging knowledge bases (KBs), such as DBpedia and Freebase, as underlying semantic model and make use of semantic annotations that contain vital cues for matching the specific information needs of users.

There is a large body of work that automatically analyzes documents and the analysis results, such as part-of-speech tags, syntactic parses, word senses, named entity and relation information, are leveraged to improve the search performance. A study [1] investigates the impact of named entity and relation recognition on search performance. However, this kind of work is based on natural language processing (NLP) techniques to extract linguistic information from documents, where the rich semantic data on the Web has not been utilized. In [2], an ontology-based scheme for semi-automatic annotation of documents and a retrieval system is presented, where the ranking is based on an adaptation of the traditional vector space model taking into account adapted TF-IDF weights.

This work can be dedicated to research in this area. Nevertheless, it provides a significantly new search paradigm. The main contributions include: (1) The rich semantics in KBs are used to yield the semantic representations of documents and queries. Based on the various semantic interpretations of queries, users can refine them to match their intents. (2) Given our emphasize on semantics of entities and relations, we introduce a novel scoring mechanism to influence document ranking through manual selection of entities and weighting of relations by users. (3) Another important feature is the support of cross-linguality, which is crucial when queries and documents are in different languages.

2 Document Retrieval Process

In this section, we present our document retrieval process, which consists of five steps. While *lexica extraction* and *text annotation* are performed offline, *entity matching*, *query refinement* and *document ranking* are handled online based on the index generated by offline processing.

Lexica Extraction. In this step, we constructed the cross-lingual lexica by exploiting the multilingual Wikipedia to extract the cross-lingual groundings of entities in KBs, also called *surface forms*, i.e., words and phrases in different languages that can be used to refer to entities [3]. Besides the extracted surface forms, we also exploit statistics of the cross-lingual groundings to measure the association strength between the surface forms and the referent entities.

Text Annotation. The next step is performed to enrich documents with entities in KBs to help to bridge the ambiguity of natural language text and precise formal semantics captured by KBs as well as to transform documents in different languages into a language independent representation. For this purpose, we employ our cross-lingual semantic annotation system [4] and the resulting annotated documents are indexed to make them searchable with KB entities.

Entity Matching. Our online search process starts with the keyword query in a specific language. Instead of retrieving documents, our approach first finds entities from KBs matching the query based on the index constructed in the lexica extraction step. These entities represent different semantic interpretations of the query and thus are employed in the following steps to help users to refine the search and influence document ranking according to their intents.

Query Refinement. Different interpretations of the query are presented for users to select the intended ones. Since interpretations correspond to entities in this step, users can choose the intended entity for refinement of their information needs. We also enable users to adjust the weights of entity relations to influence the document ranking for a personalized document retrieval. For this, the chosen entity is shown and extended with relations to other entities retrieved from KBs.

Document Ranking. After query refinement by users, the documents in different languages containing the chosen entity are retrieved from the index constructed by text annotation. Then, we exploit the semantics of entities and relations for ranking. We observe that annotated documents generally share the following *structure pattern*: every document is linked to a set of entities, where

a subset (several subsets) of these entities are connected via relations in the KB, forming a graph (graphs). In this regard, a document can be conceived as a graph containing several connected components. Leveraging this pattern, we propose a novel ranking scheme based on the focus on the chosen entity and the relevance to the weighted relations.

Focus-Based Ranking: Intuitively, given two documents d_1 and d_2 retrieved for the chosen entity 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 containing e is larger than that of d_2. Based on this rationale, we propose $\textsc{Score}_{Focus}(d, e)$ between document d and entity e to capture the focus of d on e as follows:

$$\textsc{Score}_{Focus}(d, e) = |LCC_d^e| \tag{1}$$

where LCC_d^e is the largest connected component of d containing e and $|LCC_d^e|$ represents the number of entities in LCC_d^e.

Relation-Based Ranking: Given the chosen entity e, the users can weight both the existence and the occurrence frequency of its relations to influence the document ranking. This differentiation separates the one scenario where users are interested in obtaining more detailed information about the relationship (qualitative information) from the other, where users are interested in the quantity. Let R_e be the set of relations of chosen entity e. We define $x_r = 1$ if $r \in R_e$, otherwise 0, and $y_r = \frac{|r_d|}{\log(avg_r)}$, where $|r_d|$ denotes the occurrence frequency of r in d and avg_r is the average occurrence frequency of r. Then, we propose $\textsc{Score}_{Relation}(d, e)$ between document d and entity e to capture the relevance of d to the weighted relations in R_e as follows:

$$\textsc{Score}_{Relation}(d, e) = \sum_{r \in R_e} x_r \cdot w_r^{existence} + y_r \cdot w_r^{frequency} \tag{2}$$

where $w_r^{existence}$ and $w_r^{frequency}$ are weights given by users for the existence and the occurrence frequency of relation r, respectively.

By taking into account both focus-based and relation-based ranking, we present the final function for scoring the documents as given in Eq 3.

$$\textsc{Score}(d, e) = \frac{\textsc{Score}_{Focus}(d, e) \cdot \textsc{Score}_{Relation}(d, e)}{ndl_d^e} \tag{3}$$

where ndl_d^e is the normalized document length of d w.r.t. annotations, i.e. the number of entities contained in d, which is used to penalize documents in accordance with their lengths because a document containing more entities has a higher likelihood to be retrieved. The effect of this component is similar to that of normalized document length w.r.t. terms in IR. We can compute it as

$$ndl_d^e = (1 - s) + s \cdot \frac{ef_d}{avg_{ef}} \tag{4}$$

where ef_d denotes the total number of entities in d, avg_{ef} is the average number of entities in the document collection, and s is a parameter taken from IR literature, which has been typically set to 0.2.

3 Evaluation

We now discuss our preliminary evaluation results. In the experiment, we use DBpedia [5] as the KB and Reuters Corpus Volume 1 (RCV1) as the document corpus containing about 810,000 English news articles. To assess the effectiveness of our approach, we investigate the normalized discounted cumulative gain (nDCG) measure of the top-k results instead of the common measures like precision and recall, which are not suitable to our scenario because the results can be different in relevance for each query and differ for each facet or weight used. We asked volunteers to provide keyword queries in Chinese (17 in total) along with descriptions of the intents used to set the weight for the relations, which yield the average nDCG of 0.87 and the average number of results of 612.

4 Conclusions and Future Work

In this paper, we show that the semantics captured in KBs can be exploited to allow the information needs to be specified and addressed on the semantic level, resulting in the semantic representations of documents and queries, which are language independent. The user feedback on our demo system [6] suggests that the proposed approach enables *more precise refinement* of the queries and is also valuable in terms of the *cross-linguality*. In the future, we plan to advance the query capability to support keyword queries involving several entities and conduct more comprehensive experiments to evaluate our system.

Acknowledgments. This work is supported by the European Community's Seventh Framework Programme FP7-ICT-2011-7 (XLike, Grant 288342) and FP7-ICT-2013-10 (XLiMe, Grant 611346). It is also partially supported by the German Federal Ministry of Education and Research (BMBF) within the SyncTech project (Grant 02PJ1002) and the Software-Campus project "SUITE" (Grant 01IS12051).

References

1. Chu-Carroll, J., Prager, J.M.: An experimental study of the impact of information extraction accuracy on semantic search performance. In: CIKM. (2007) 505–514
2. Castells, P., Fernández, M., Vallet, D.: An adaptation of the vector-space model for ontology-based information retrieval. IEEE Trans. Knowl. Data Eng. **19**(2) (2007) 261–272
3. Zhang, L., Färber, M., Rettinger, A.: xlid-lexica: Cross-lingual linked data lexica. In: LREC. (2014) 2101–2105
4. Zhang, L., Rettinger, A.: X-lisa: Cross-lingual semantic annotation. PVLDB **7**(13) (2014) 1693–1696
5. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellmann, S.: DBpedia - A crystallization point for the Web of Data. J. Web Sem. **7**(3) (2009) 154–165
6. Färber, M., Zhang, L., Rettinger, A.: Kuphi - an investigation tool for searching for and via semantic relations. In: ESWC. (2014)