Linked Data Annotated Document Retrieval

Claudia Exeler, Jörg Waitelonis, and Harald Sack

Hasso-Plattner-Institute for IT-Systems Engineering, Prof.-Dr.-Helmert Str. 2-3,
14482 Potsdam, Germany
claudia.exeler@student.hpi.de, joerg.waitelonis@hpi.de,
harald.sack@hpi.de

1 Introduction

Search engines traditionally suffer drawbacks from ambiguities of natural language, which users often solve via query refinement. In contrast to web search, querying document collections of limited size (e.g. blogs, multimedia collections, or libraries) can quickly lead to empty result sets because the wrong choice of keywords may eliminate the only relevant document. Besides query expansion [1], a solution to overcome these shortcomings is to explicitly map the document contents to knowledge bases, and to exploit provided information to take into account semantic similarity and relatedness among documents and queries. For this purpose, we have developed and evaluated a *Semantic Search* system, which combines traditional keyword-based search with Linked Open Data knowledge bases, in particular DBpedia. We present two novel retrieval approaches and contribute an evaluation data set with semantically annotated documents, search queries, as well as relevance judgements.

2 Linked Data Annotated Document Retrieval

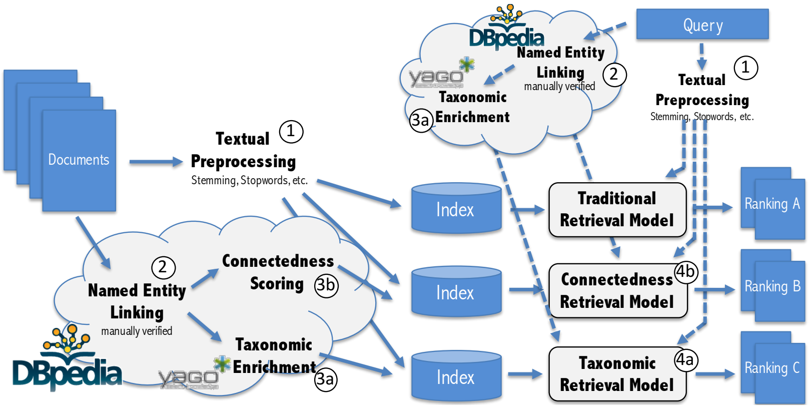

Fig. 1. Evaluation architecture overview.

The search process of our Semantic Search system is depicted in Fig. 1. It first preprocesses documents and queries using state-of-the-art indexing methods (1). Then, *Named Entity Linking* annotates document content and queries in terms of natural language text with DBpedia entities (2). The DBpedia IRIs become part of the index and are used to create a Generalized Vector Space Model (GVSM), where index terms are not considered pairwise orthogonal. Inspired by the semantic GVSM of [3], we propose two extensions: The *taxonomic approach* (3a & 4a), which determines term correlations based on YAGO [1] classes, as well as *connectedness weighting* (3b & 4b). Together with the traditional keyword based search, which is used as baseline, the system generates three indexes for the different retrieval models.

With the goal to increase recall, the taxonomic approach determines documents containing entities that are not explicitly mentioned in, but strongly related to the query, e. g. if a document entity belongs to the same class as a query entity. We exploit these relationships to also identify documents that can serve as helpful recommendations if none or only few directly relevant documents exist, which is a frequent scenario when searching on limited document collections. The term vectors $\boldsymbol{t_i}$ are constructed from the entity vector $\boldsymbol{e_i}$ of the entity it represents and the set of classes $c(e_i)$ the entity is member of:

$$\boldsymbol{t_i} = \alpha_e \boldsymbol{e_i} + \alpha_c \frac{\boldsymbol{v_i}}{|\boldsymbol{v_i}|}, \text{ with } \boldsymbol{v_i} = \sum_{c_j \in c(e_i)} w(c_j, e_i) \times \boldsymbol{c_j}. \tag{1}$$

The $\boldsymbol{e_i}$ and $\boldsymbol{c_j}$ are pairwise orthogonal vectors with n dimensions, where each dimension stands for either an entity or a class. Since not every shared class means the same level of relatedness between two entities, not all classes should contribute equally strong to the similarity score. Assigning weights $w(c_j, e_i)$ that express the relevance of the class c_j to the entity e_i achieves this effect. Traditional semantic similarity measures are most suited for this purpose if they consider class specificity, so we have used the measure proposed by Resnik [2]. The factors α_e and α_c define the contribution of entities compared to their classes and should incorporate normalization to keep unit length for the term vectors. With larger α_e, a document with few occurrences of the queried entity will be preferred over a document with frequent occurrence of related entities. The text index integrates into this model by appending the traditional document vector to the entity-based document vectors.

The second approach addresses the issue that term frequency is not always the most appropriate indicator of the relevance of a term (or entity) within a document. We propose to use *connectedness weighting* instead, which defines an entity's relevance within a document based on how strongly it is connected within the *document subgraph* D. D includes all entities that are linked within the document as well as all entities from the knowledge base that connect at least two entities from the document. To obtain an undirected graph as required for connectedness calculation, the function $rel(e_i, e_j)$ is applied. It returns true

[1] `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/`

iff there exists a relation from e_i to e_j or from e_j to e_i. Each entity $e_i \in D$ has a set E_i of directly connected entities and a set F_i of indirectly connected entities:

$$E_i = \{e \in D | rel(e, e_i)\} \text{ and } F_i = \{e \in D | \exists x : rel(e, x) \wedge rel(x, e_i)\} \qquad (2)$$

Based on these sets, connectedness is calculated as follows:

$$cn(e_i, d) = 1 + (|E_i| + |F_i|) \times \frac{|D|}{n_d}, \text{ where } n_d = \sum_{e_j \in D} |E_j| + |F_j|. \qquad (3)$$

Entities may have no connections to any other entities in the document subgraph. Since they are nevertheless relevant to the document, we add 1 to all scores. The score is normalized by the average number of connected entities over all $e \in D$ ($|D|/n_d$) to create comparability between different documents. This is otherwise lacking because entities are more likely to be connected to other entities in documents with more annotations. In addition, a single connection to another entity is more significant in a sparse document subgraph than in a dense one.

Whether or not a word or entity has a large power to distinguish relevant from non-relevant documents depends on the corpus. Therefore, we keep the traditional Inverse Document Frequency (IDF) to calculate a term's distinctness. The entity vectors' values are thus cn-idf values, i.e. $w(e_i, d) = cn(e_i, d) \times idf(e_i)$.

3 Evaluation

There are currently no datasets that provide semantically annotated documents *and* queries with relevance assessments, so we have compiled a new dataset from 331 texts. They have an average length of 570 words, with 3 to 255 manually revised annotations. We also assembled and manually annotated a set of 35 queries.

For every query, the top 10 ranked documents from text-, class- , and connectedness search were presented to users in random order. The users were asked to assign every document to one of the five categories based on its relation to the query: Document is relevant (corresponding to a score of 5), parts are relevant (3), document is related (3), parts are related (1), irrelevant (0). The rounded arithmetic mean of all users' scores determines the relevance score in the ground truth. In total, 64 users have participated in the relevance assessments. All queries have been assessed by at least 8 participants[2].

Tab. 1 shows that the inclusion of semantic annotations and similarities clearly improves retrieval performance compared to the text search baseline. The taxonomic approach not only increases recall, but also improves the ranking quality, measured by Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG).

The connectedness approach performs better than text search, but worse than the other semantic methods, including the simple "Concept+Text" approach,

[2] The ground truth dataset is published at: `http://s16a.org/node/14`

Table 1. Evaluation Results

Method	MAP	NDCG	MAP@10	NDCG@10	Recipr. Rank	Prec@1
Text (baseline)	0.696	0.848	0.555	0.743	0.960	0.943
Concept + Text	0.736	0.872	0.573	0.761	0.979	**0.971**
Connectedness (only)	0.711	0.862	0.567	0.752	**0.981**	**0.971**
Connectedness (with tf)	0.749	0.874	0.583	**0.766**	0.979	0.943
Taxonomic (no similarity)	0.766	0.875	0.603	0.758	0.961	0.943
Taxonomic (Resnik-Zhou)	**0.768**	**0.877**	**0.605**	0.762	0.961	0.943

where entities are treated as regular index terms within Lucene's default model. This is surprising because when we let the users directly compare the rankings produced by the three approaches, connectedness performed best. The evaluators had to identify the best (2.0) and second-best (1.0) rankings, which resulted in an average score of 1.09 for connectedness, followed by 1.01 for the taxonomic approach and 0.90 for the baseline. This seeming contradiction hints at a difference between information retrieval evaluation measures and user perception of ranking quality. The evaluators seem to have judged mainly by the very top few documents. Connectedness outperforms the other approaches in this respect, as shown by the reciprocal rank and precision@1 (Tab. 1). Also, connectedness performs best when related documents are not considered relevant. Combining the connectedness measure with tf weights leads to clear improvements.

4 Conclusions and Future Work

Both proposed methods seem to achieve improvements over traditional text retrieval. Open questions, which are to be answered in future work, include how well the models would perform with other knowledge bases (e.g. Wikidata), what other semantic relations between entities are valuable for document retrieval, and how the semantic similarity can be given more influence. The annotation of queries with classes may improve the retrieval, and so could the combination of the two proposed methods. Furthermore, the main ideas could be transferred to an adapted Language or Probabilistic Retrieval Model.

References

1. V. M. Ngo and T. H. Cao. Ontology-based query expansion with latently related named entities for semantic text search. *Advances in Intelligent Information and Database Systems*, 2010.
2. P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th Int. Joint Conference on Artificial Intelligence*, pages 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
3. G. Tsatsaronis and V. Panagiotopoulou. A generalized vector space model for text retrieval based on semantic relatedness. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics Student Research Workshop on EACL 09*, (April):70–78, 2009.