Ontology-Supported Document Ranking for Novelty Search

Michael Färber*

Karlsruhe Institute of Technology (KIT), 76131 Karlsruhe, Germany
michael.faerber@kit.edu

Abstract. Within specific domains, users generally face the challenge to populate an ontology according to their needs. Especially in case of novelty detection and forecast, the user wants to integrate novel information contained in natural text documents into his/her own ontology in order to utilise the knowledge base in a further step. In this paper, a semantic document ranking approach is proposed which serves as a prerequisite for ontology population. By using the underlying ontology for both query generation and document ranking, query and ranking are structured and, therefore, promise to provide a better ranking in terms of relevance and novelty than without using semantics.

Keywords: Document ranking, Ontology-based information extraction, Novelty detection, Semantic similarity.

1 Motivation

The existence and steady growth of the Web has granted us vast amounts of web documents in which contained information can be discovered and utilised for certain information needs. Some of the existing information extraction (IE) techniques make use of background information provided by Semantic Web ontologies. In the past, various ontology-based information extraction (OBIE) systems have been proposed, where ontologies are used within the IE process. Although there exist quite a lot of notable ontologies, in many application areas appropriate ontologies are, due to domain-specificity, too small and, hence, need to be populated in terms of adding instances and properties. For ontology population, it is a crucial task to find new textual information which is relevant to the domain expert, but has not been stored in the knowledge base (KB) and, therefore, has been made usable. In this work, we focus on the worthwhile interplay between an existing KB and a text document corpus, which – in case of the use case of trend detection – is created on demand.

Within the area of ontology population, we propose a novel approach for document ranking in the context of structural search for "novel" items in text documents. We claim that semantics can be used to rank documents according to their expected novel items contained therein.

* Work leading to this paper has been partially supported by the German Ministry of Education and Research (BMBF) under grant no. 02PJ1000.

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7882, pp. 639–644, 2013.

2 Related Work

Our approach is part of an ontology population system with the task of finding relevant and novel information and integrating it into a – e.g., company wide – KB. There are already many OBIE systems [1]. However, concerning novelty search on documents, current approaches show only little [2] or no semantic components [3, 4], although semantics can resolve inconsistencies and ambiguities. Existing approaches are subject to different definitions of novelty and different application areas and granularities. Within the TREC "novelty track" in 2002–2004 [5], systems for detecting novelty were designed. However, the task took place on sentence level, was limited to event and opinion detection, and was aligned for non-domain-specific texts such as news. A similar case is the novelty detection task of the Text Analysis Conference (TAC) Knowledge Base Population (KBP) track [6]. Li and Croft [2] address the field of novelty formalisation in depth. Under the semantic point of view, they merely make use of a low-key named entity recognition and classification (NERC) component and primarily rely on statistical patterns. Zhang et al. [4] regard the challenge of novelty and redundancy detection as a filtering process. Documents are filtered at first according to relevance to the topic, and in a second step according to novelty defined as non-redundancy with respect to previously seen documents. Contrary to systems like "Newsjunkie" [3], we face domain-specific documents like technical reports and patents, and therefore do not have to deal with the problem of analysis of huge amounts of articles in a very short time period, known as "burst of novelty".

Besides the novelty aspect, our work touches upon the research area of query generation as well as graph comparison techniques and similarity metrics. Work here [7–10] might show good results for query suggestion or expansion techniques. Our novel approach, however, uses an underlying ontology as a bridge for both query generation and document ranking.

Last but not least, Aleman-Meza et al. [11] and several researchers at the TAC KBP track [6] whose task it was to find property values in documents (called "slot filling") provide a document ranking approach which also exploits named entities (NEs) found in documents. In the first case, a weighting schema is proposed, where domain experts need to assign weights to the edges between classes of the KB schema in order to model the relatedness. The existence of huge ontologies like DBLP and many different data sources is assumed here. Contrary to this assumption, we want to populate our own, rather small, domain-specific ontology with instances and properties and need to take novelty detection into consideration.

3 Proposed Approach

Given our own KB with instances and schema, our goal is to search for documents and to rank them, so that the documents most novel to the KB and relevant to the query and to the KB have the highest ranking. In the overall OBIE system

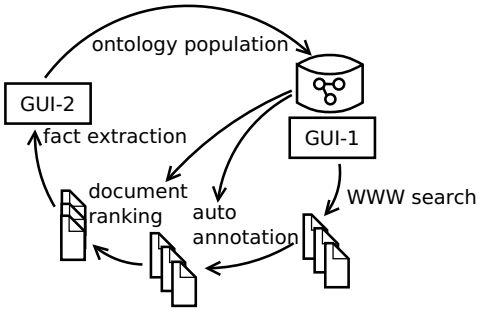

Fig. 1. According to a user's context a structured query is generated with the help of an underlying ontology. After the creation of a document corpus (using the query in an unstructured fashion), annotation and ranking of documents are performed. In a further step, on which we do not focus on, the annotations are verified by the user and used for populating the ontology. In succeeding search rounds search is based on the enriched ontology.

in which our approach is embedded, a second step follows in which the user is able to import phrases marked in the document into his/her KB as property values. Figure 1 gives an overview of the interplay between an ontology and document texts with potentially novel information. In the following, we describe our ranking approach.

Assuming that we have a pre-defined KB schema with assigned weights on the edges expressing the strengh of relatedness and some instances, we start our search by defining the search query – and, hence, the query graph – by the user and his/her context. Besides instances and property values from the KB, additional search keywords can be defined by the user. After expanding the query graph with neighbouring entities of the KB (or neighbouring instances of merely the targeted entity type), we can transform the query graph plus additional keywords into a keyword phrase for simple document search, getting a crawled set of web documents. Of course, we can also operate on a fixed document collection, although this would hamper the overall goal of getting external novel information like in the use case of trend detection and forecast.

As the extended query graph is a subgraph of the KB instance graph and each instance has a fixed set of possible properties, we can find out which relationships (i) between instances and property values and (ii) between instances and other instances of the KB exist and which are still missing. To include the "real" filling degree in terms of personalised importance or novelty degree, we use the weights of the edges in the KB schema graph. By means of the KB, we construct for each document a graph containing all instances found as NEs in the focused document and their relationships among these instances read from the KB. According to further features such as the frequency of the found NEs, additional weights can be assigned to the nodes in the document graphs. For each document,

we can compute a final score compliant with the local severity of found instances in the document, with their novelty degree (inverse filling degree), and with the actual weighting of edges in the KB schema graph. New detected NEs and string matches are also included.

The documents are ranked according to the document scores they obtained. Furthermore, we can use implicit user feedback in the following way: If the user determines which properties or instances are important and novel in the focused document, the weights in the KB schema graph between the classes of the instances (or properties) which were found in the document are adapted. By this means, we can defer to the personal views what relationships between certain classes and properties (or other classes) are of great significance.

The proposed ranking approach is geared to the need of having an approach for ranking documents as a prerequisite for the ontology population task. This involves the inclusion of the novelty aspect into ranking and the adaptation of context and user-dependent association weights between classes.

4 Implementation and Research Methodology

The proposed framework of ontology supported novelty search is currently under development, so that experiments and evaluation could not be performed yet. As use case we chose technology companies, since they are interested in technology forecasts and novelty detection. The lightweight use case ontology consists of classes like technology, company, product, and person. For a valid and comparable evaluation, we plan to evaluate our approach also on a non-specific domain, using the AQUAINT collection, which consists of newswire articles, as used in the TREC 2005 HARD track. Here, DBpedia will be used as underlying KB.

Annotation is done by the wikify service of the Wikipedia Miner [12]. We adopt ideas from wikifier, but adapt it to specific domains, by using the content of our domain-specific semantic-based wiki. In order to detect also new entities, property values, and relationships, we use GATE[1], a well-established rule-based framework.

Our research focuses on semantic document ranking. We implement and plan to evaluate a ranking score function as proposed above. Concerning our domain-specific use case, the final evaluation will be done by students and experts in companies. During the evaluation, we compare the approach of manually assigning weights to the edges in the schema graph with the approach of learning weights. Possible evaluation scenarios entail: 1. We measure whether the users need less time to find a specific amount of relevant and novel documents in comparison to the time they needed in case of using generic search engines like Google. 2. We can also determine whether more relevant and novel documents were found in a specific time interval. This is the main aim of innovation partners in companies and serves as practical motivation.

[1] `http://gate.ac.uk`

5 Conclusion and Prospects

Semantic-based solutions for document ranking do not regard novelty as a criterium so far. In this work, a new ranking approach is proposed. It is designed to improve document retrieval, since users generally face the problem of being commited to review too many text documents containing irrelevant or already known information. With the help of the proposed ranking schema, the more relevant and potentially novel information a document contains, the higher it is ranked and, hence, more likely to be worth reading and the more useful for ontology population. The next steps will involve the implementation and valid evaluation of the semantic ranking approach. In the medium term, we plan to integrate our work into a theoretical foundation like Markov random models.

References

1. Wimalasuriya, D.C., Dou, D.: Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science 36(3), 306–323 (2010)
2. Li, X., Croft, W.B.: An information-pattern-based approach to novelty detection. Information Processing & Management 44(3), 1159–1188 (2008)
3. Gabrilovich, E., Dumais, S., Horvitz, E.: Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In: Proceedings of the 13th International Conference on World Wide Web, WWW 2004, pp. 482–490. ACM, New York (2004)
4. Zhang, Y., Callan, J., Minka, T.: Novelty and redundancy detection in adaptive filtering. In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2002, pp. 81–88. ACM, New York (2002)
5. Soboroff, I., Harman, D.: Novelty detection: the TREC experience. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, HLT 2005, pp. 105–112. Association for Computational Linguistics, Stroudsburg (2005)
6. Ji, H., Grishman, R., Dang, H.T.: Overview of the TAC2011 Knowledge Base Population Track (2011)
7. Bendersky, M., Metzler, D., Croft, W.B.: Effective query formulation with multiple information sources. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM 2012, pp. 443–452. ACM, New York (2012)
8. Meij, E., Bron, M., Hollink, L., Huurnink, B., de Rijke, M.: Learning Semantic Query Suggestions. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 424–440. Springer, Heidelberg (2009)
9. Bendersky, M., Croft, W.B.: Modeling higher-order term dependencies in information retrieval using query hypergraphs. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2012, pp. 941–950. ACM, New York (2012)

10. Bendersky, M., Metzler, D., Croft, W.B.: Parameterized concept weighting in verbose queries. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, pp. 605–614. ACM, New York (2011)
11. Aleman-Meza, B., Arpinar, I.B., Nural, M.V., Sheth, A.P.: Ranking Documents Semantically Using Ontological Relationships. In: Proceedings of the 2010 IEEE Fourth International Conference on Semantic Computing, ICSC 2010, pp. 299–304. IEEE Computer Society, Washington, DC (2010)
12. Milne, D., Witten, I.H.: Learning to link with wikipedia. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, CIKM 2008, pp. 509–518. ACM, New York (2008)