A semantic wiki for novelty search on documents

Michael Färber*
Karlsruhe Institute of Technology (KIT)
Institute AIFB
76131 Karlsruhe
michael.faerber@kit.edu

Achim Rettinger
Karlsruhe Institute of Technology (KIT)
Institute AIFB
76131 Karlsruhe
rettinger@kit.edu

ABSTRACT

Technology-oriented companies are typically interested in monitoring developments concerning their technologies. However, most companies, especially SMEs, don't have an efficient process how this is achieved. If at all, efforts are mostly limited to uncoordinated keyword queries on web resources. Here, we present a semi-automatic approach that allows for structured and continuous detection of relevant, novel and domain specific documents appearing on the Web. Our system is based on a semantic wiki where the domain expert is able (i) to store all relevant information in an adequate knowledge base with the ability for monitoring and trend mining and (ii) to import detected novel items such as future technologies and their properties to the knowledge base in a continuos fashion. The latter is achieved by generating a structured query based on the user context and by representing found documents as semantic graphs. In this way, novel items can be found easier and in a semi-automatic fashion.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Economics

Keywords

semantic wiki, novelty detection, document ranking, ontology-supported information extraction.

1. MOTIVATION

Technology forecast and trend detection are indispendable tasks for technology companies in order to be informed about market developments and inventions in their fields. With the advent of more and more documents on the Web, companies face the task of extracting relevant and novel information for this purpose. Currently, this has to be done usually purely manually and without any structured background data making it a very time-consuming task. Therefore, we provide a semi-automatic process for trend detection and monitoring services. We present a semantic wiki-based application which is based on ontology-based information extraction (OBIE) where ontologies are used within the information extraction (IE) process. Since usually appropriate ontologies regarding technologies and their properties are missing or are too small, we focus our work on the crucial task of how to efficiently 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 some sense.

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

DIR 2013, April 26, 2013, Delft, The Netherlands.

2. RELATED WORK

Within the TREC "novelty track" in 2002–2004 [2], 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. *Newsjunkie* [1] is also geared to detecting novelty by comparing a new document against an existing document collection. Contrary to such systems, we face domain-specific documents like technical reports and patents, and therefore do not have to deal with the problem of analysing huge amounts of articles in a very short time period, known as "burst of novelty". Instead of purely statistical measures, our approach is based on semantic technologies.

3. DOCUMENT RANKING AND ONTOLOGY POPULATION

Figure 1 gives an overview of the interplay between an ontology and documents with potentially novel information: 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 both the query and the KB have the highest ranking. In a second step the user is able to import phrases marked in the document into his/her KB as property values.

Concerning the first part, Semantic MediaWiki[1] as an instance of a semantic wiki is assigned the central role: The user is able to create new wiki pages (within the semi-automatic process or just manually) and to add

[1]http://semantic-mediawiki.org/

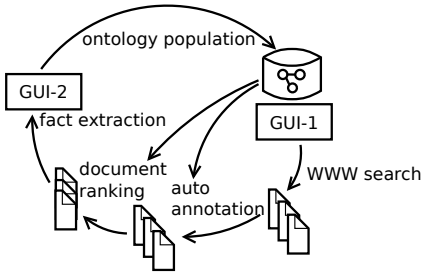


Figure 1: According to a user's context a structured query is generated with the help of an underlying ontology. Afterwards, ranking is performed using annotated document corpus. In the last step, annotations are verified by the user and used for populating the ontology. In succeeding search rounds search is based on the enriched ontology.

appropriate properties with the help of a class-specific form (see figure 2). Internally, all data is stored in a structured way. The wiki allows the user to create a search query out of the context by taking instances and property values (from the KB) as well as search keywords written by the user. After an optional expanding of the query graph with neighbouring entities, we can generate the final query graph. Since all documents are annotated with the help of named entity recognition tools[2], we can compare the generated query graph with all document entity graphs (generated from extracted named entities). Ranking of the documents is facilitated by weights which were assigned to every relation in the KB schema graph. We can use implicit user feedback in the following way: If a user imports some novel item as a new property or instance, the weights in the KB schema graph 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 and should be reinforced for next search sessions.

Our focused use cases are determined by our use case partners[3] which are medium-sized technology companies. Hence, the lightweight ontologies we used consist of classes like *technology*, *institution*, and *product*. As document corpus, web documents retrieved by search engine requests are considered. In addition, trend detection in conjunction with patents can be enabled by using the patent database Espacenet[4], where access to over 70 million patent documents and their meta data is provided.

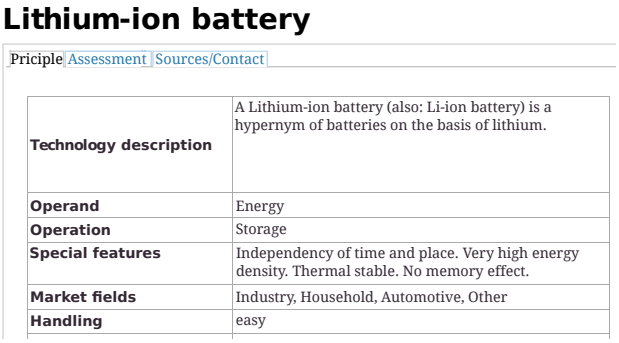
4. CONCLUSION

Existing processes and tools for trend mining and technology watch are often only rudimentary implemented, especially in SMEs. We have presented a semantic wiki for storing and displaying structured information about a specific

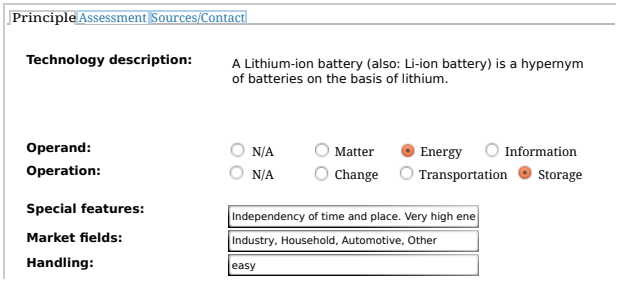
[2]One of these tools is the wikify service of the Wikipedia Miner (http://wikipedia-miner.cms.waikato.ac.nz/) which we adapt by using the content of our domain specific semantic-based wiki. In order to detect also new entities, property values, and relationships, we use GATE (http://gate.ac.uk), a well-established rule-based framework.

[3]Industry partners within the German research project *syncTech* (http://synctech-innovation.de).

[4]http://worldwide.espacenet.com



(a)



(b)

Figure 2: Screenshots of a Semantic MediaWiki: (a) displaying technology property values within a wiki page (b) edit functionality using form.

domain (industrial technology field) and for generating a context-aware semantic search query. With the help of a new 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 used for ontology population. Due to the use of structured information and approriate background data the way of doing trend mining can be changed towards a semi-automatic process with better search and monitoring capabilities.

5. REFERENCES

[1] Evgeniy Gabrilovich, Susan Dumais, and Eric Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 482–490, New York, NY, USA, 2004. ACM.

[2] Ian Soboroff and Donna Harman. Novelty detection: the TREC experience. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 105–112, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.