Towards a Semantically Enriched Online Newspaper

Ricardo Kawase, Eelco Herder, Patrick Siehndel

L3S Research Center, Leibniz University Hannover, Germany
{kawase, herder, siehndel}@L3S.de

Abstract. The Internet plays a major role as a source of news. Many publishers offer online versions of their newspapers to paying customers. Online newspapers bear more similarity with traditional print papers than with regular news sites. In a close collaboration with Mediengruppe Madsack - publisher of newspapers in several German federal states, we aim at providing a semantically enriched online newspaper. News articles are annotated with relevant entities - places, persons and organizations. These annotations form the basis for an entity-based 'Theme Radar', a dashboard for monitoring articles related to the users' explicitly indicated and inferred interests.

1 Introduction

Traditional print media are nowadays replaced or complemented by online media. Most publishers of international, national and local newspapers use the Web as an additional communication channel. Many news sites that are connected to a print newspaper also offer online subscriptions or provide content as pay-per-view. A commonly used solution is subscription-based access to an online newspaper, which is a digital copy of the print newspaper, often with additional features for search, recommendation or archiving. However, in most cases, these additional features are based on content analysis, manual interlinking by the editors and collaborative filtering. In this paper, we present our work towards an semantically enriched online newspaper, which is a currently running collaboration between the L3S Research Center and Madsack GmbH & Co. KG.

1.1 Madsack

Madsack GmbH & Co. KG is a German media group with headquarters in Hannover, Germany. Its core business comprises the production of several regional newspapers in Lower Saxony, Schleswig-Holstein, Mecklenburg-Western Pomerania, Saxony, Hessen and Saxony-Anhalt. Madsack is the sixth largest publishing house in Germany[1]; in the year 2012, the average circulation of their 18 paid newspapers amounted to 939,590 copies[2].

[1] http://www.media-perspektiven.de/uploads/tx_mppublications/05-2010_Roeper.pdf

[2] http://www.madsack.de/fileadmin/content/downloads/Geschaeftsbericht_MGM_2012-2013_web.pdf

The digital business of Madsack media group includes the distribution of editorial content (e-paper, mobile apps, usually using the brand of the corresponding daily newspapers), marketing services (e.g. programming of websites and apps) as well as collaborations with online marketplaces.

We focus on Madsack's e-paper product. The e-paper is a Web environment that allows subscribers to access the daily editions of the newspaper in digital format. The environment is restricted to paying subscribers, who are required to log in with their personal username and password. Once they are logged in, the website presents the reader current daily newspaper editions. The online newspaper holds the same design as the printed version. Every morning, except on Sundays (there are no editions printed on Sundays), a news daily edition is available on the website.

2 Enrichment

In order to effectively archive, categorize and publish news articles, most larger media companies have documentation departments that assign labels, categories or terms to news articles. Due to the increasingly large amount of items and the need for the term assignment to be quick [3], automatic semantic annotation is increasingly considered as an alternative for human annotation. Several established off-the-shelf tools for knowledge extraction and semantic annotation are readily available, including DBpedia Spotlight [4], AIDA [7], Open Calais, Wikimeta and Zemanta [2]. Wikipedia Miner [6] directly makes use of the evolving Wikipedia structure; the toolkit includes search and annotation services and provides links to relevant Wikipedia articles. In a comparison of entity-annotation systems [1], Wikipedia Miner consistently scored high in terms of recall and F1.

We semantically enriched the news articles by identifying entities and types. For this purpose, we use the Wikipedia Miner[5] service as an annotation tool. First, detected words are disambiguated using machine learning algorithms that take the context of the word into account. This step is followed by the detection of links to Wikipedia articles (which will later be aligned with DBpedia entities). By using a predefined threshold, we ensured that only those words that are relevant for the whole document are linked to articles. The goal of the whole process is to annotate a given news article in the same way as a human would link a Wikipedia article. We set up a local deployment of Wikipedia Miner and trained the models on top of a German Wikipedia dump from February, 2014[3].

After annotating the content of the news articles, with the identified entities in hand, we query DBpedia in order to gather further information regarding the entities. Specifically, we explore their relationships through the predicate **rdf:type** to extract the type of the entity given by DBpedia's ontology (dbpedia-owl). Although several different types are identified, we selected the three most relevant types for news articles: **dbpedia-owl:Place**, **dbpedia-owl:Person** and **dbpedia-owl:Organisation**. These three types were reported by Madsack's editorial staff to be the most relevant for their readers. Additionally, as we describe in Section 3, it is important to avoid an overload of features and information to the readers. Thus, we aim at having just a few and very useful facets that can improve relevant news retrieval.

[3] `http://dumps.wikimedia.org/dewiki/20140216/`

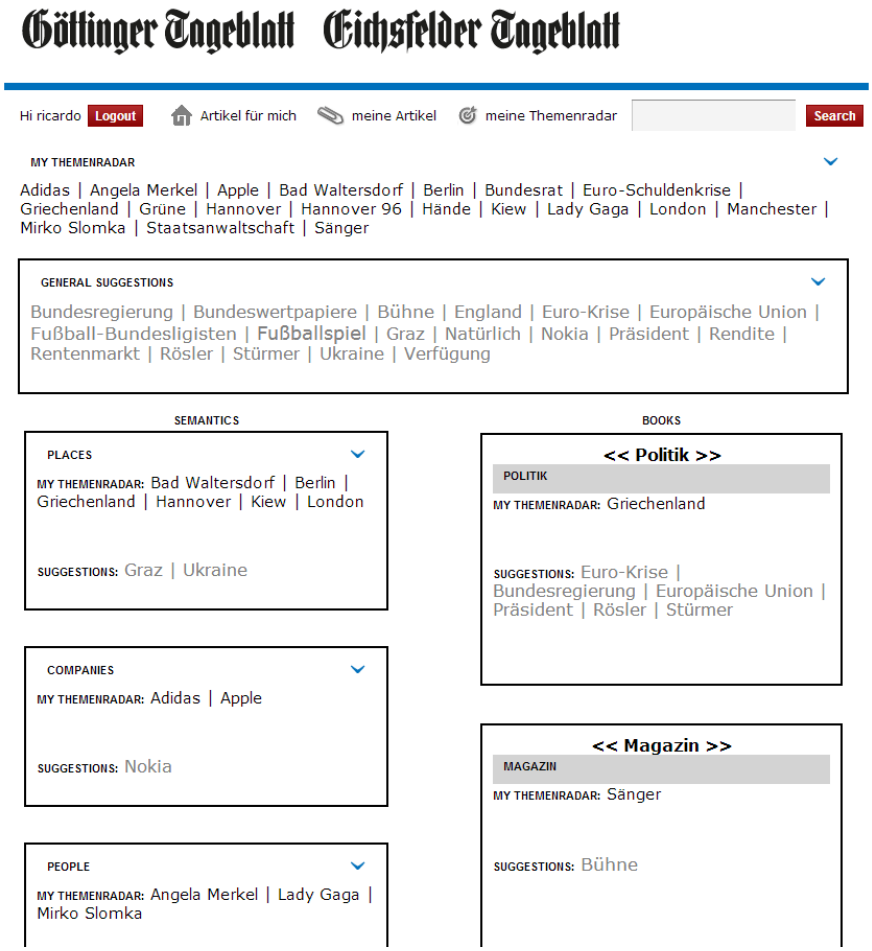

Fig. 1. Madsack's e-paper prototype interface.

3 User Interface

From the end users' (the readers') perspective, the main innovation of the e-paper is the so called 'Themenradar' (Theme Radar). The Theme Radar provides users with shortcuts to news articles that pertain to their particular interests - as explicitly indicated by subscribing to an entity (which represents a theme or topic), combined with the entities that most often occur in the articles that the user read so far. Augmenting the 'Theme Radar' with the assistance of semantically enriched data is, in fact, one of our main goals in this collaboration.

Figure 1 depicts the first prototype of the 'Theme Radar'. It consists of a dashboard of the readers' interests. The 'Theme Radar' works as a semantically enhanced topic dashboard that enables readers to get suggestions for themes and topics, to manage their

topics and to get personalized news recommendations based on entity co-occurrences, linked data relations and the aforementioned semantic properties types.

Based on the users' activity logs, the system automatically builds the 'Theme Radar'. Top entities of interest are presented to the users in their 'Theme Radar' with additional suggested entities and recommended articles (based on entity co-occurrence). In the interface, these entities are grouped by type and also by 'Book' (Books are sections within the newspaper, as predefined by the editors - such as 'Sports' and 'Politics'). Additionally, the users can manually add entities to their profiles, which get higher weights in the news recommendation process.

4 Conclusions

In this paper, we presented our work towards a semantically enriched online newspaper, which - to the best of our knowledge - is the first of its kind in a fully commercial setup. We are currently on a stage of interface designing which, in a commercial product, requires the validation and approval from several stakeholders. As future work, we plan to evaluate the quality of the annotations with user feedback and to perform an analysis of online reading behavior, with a focus on the semantic aspects. Building upon these steps, we plan to develop and entity-based news recommender that fulfills Madsack's online readers' interests, and to evaluate them in practice.

5 Acknowledgement

We would like to thank the Madsack Online GmbH & Co. KG team for the collaboration opportunity and the support during the implementation presented in this work.

References

1. M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 249–260. International World Wide Web Conferences Steering Committee, 2013.
2. A. Gangemi. A comparison of knowledge extraction tools for the semantic web. In *The Semantic Web: Semantics and Big Data*, pages 351–366. Springer, 2013.
3. A. L. Garrido, O. Gómez, S. Ilarri, and E. Mena. An experience developing a semantic annotation system in a media group. In *Natural Language Processing and Information Systems*, pages 333–338. Springer, 2012.
4. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.
5. D. Milne and I. H. Witten. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA, 2008. ACM.
6. D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. *Artificial Intelligence*, 194:222–239, 2013.
7. M. A. Yosef, J. Hoffart, I. Bordino, M. Spaniol, and G. Weikum. Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453, 2011.