Automatically Annotating Web Pages Using Google Rich Snippets

Frederik Hogenboom
fhogenboom@ese.eur.nl

Flavius Frasincar
frasincar@ese.eur.nl

Damir Vandic
vandic@ese.eur.nl

Jeroen van der Meer
jeroenvdmeer@gmail.com

Ferry Boon
ferry.boon@gmail.com

Uzay Kaymak
kaymak@ese.eur.nl

Econometric Institute
Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

ABSTRACT

We propose the Automatic Review Recognition and annOtation of Web pages (ARROW) framework, a framework for Web page review identification and annotation using RDFa Google Rich Snippets. The ARROW framework consists of four steps: hotspot identification, subjectivity analysis, information extraction, and page annotation. We evaluate an implementation of the framework by using various Web sites. Based on the evaluation we conclude that our framework is able to properly identify the majority of reviews, reviewed items, and review dates.

1. INTRODUCTION

Despite the technological advances of the last decades, it remains difficult for machines to understand information contained in Web pages on the World Wide Web. One of the pillars of the Semantic Web is to define the content of the Web pages semantically (i.e., as concepts with meaning) in order to make data machine understandable. The ability of computers to automatically process and interpret data will support new functionality on the Web.

Google's Rich Snippets is a service for Web page owners to add semantics to their (existing) Web page using the semantic vocabulary provided by Google. Up until now the vocabulary is rather limited in its number of concepts (Person, Review, Review Aggregate, Product, and Organization, Recipe, and Video). Future applications are promising, e.g., when searching for "Christian Dior" products, with Rich Snippets one is able to state that all results with "Christian Dior" as a person should be ignored.

For annotating Web sites built from structured data from a database, it would be sufficient to identify concepts in the generated pages and add the corresponding attributes to the Web page while generating the HTML output. Not all Web pages are built from databases and thus pre-generation of annotations is not always possible. The latter type of Web pages require manual annotation, which can be a tedious task. Hence, we present a method to automatically read and annotate Web pages, using the RDFa attributes as defined in Google Rich Snippet's vocabulary. The Automatic Review Recognition and annOtation of Web pages (ARROW) framework reads Web pages, identifies reviews, and annotates the pages with the RDFa attributes defined by Google Rich Snippets. An extended version of this paper containing more details on the framework is to be presented at the 26th ACM Symposium on Applied Computing [4].

DIR 2011 February 4, 2011, Amsterdam, The Netherlands.

2. RELATED WORK

In this paper, we focus on unsupervised Web information extraction systems, as they can be fully automated and do not require pre-annotated documents for training. Based on Web page contents, unsupervised methods try to find a pattern on the Web page, e.g., a set of recurring HTML tags or specific text strings. Examples of unsupervised Web information extraction systems are RoadRunner [2] and DeLa [5]. To identify the attributes of the reviews, e.g., author, date, etc., these systems employ unsupervised information extraction methods for Web pages. These methods can be divided into tag-based approaches, text-based approaches, and hybrid approaches. The tag-based approaches derive a wrapper for the Web site based on the structural characteristics of a Web page. Text-based approaches focus on the textual content of a Web page. Last, the hybrid approaches are a combination of the tag-based and text-based approaches and hence contain elements of both methods.

There are three different approaches to review annotation. First of all, Microformats [3] is a collection of formats that makes the representation of semi-structured information such as reviews possible. In the case of reviews, the hReview microformat can be encountered on various Web sites. Second, the W3C is working on extending the HTML language, as part of the HTML5 specification, to allow native support for annotations as described by the Microdata format. The third and final option is RDFa. RDFa extends XHTML with a set of attributes that allow the XHTML code to be enriched with metadata. Although RDFa is aimed towards extending XHTML, its attributes can also be used in HTML as most RDFa parsers will recognize these attributes.

3. ARROW FRAMEWORK

Google Rich Snippets supports a limited vocabulary of RDFa entities and their attributes. Our main focus is on

recognizing and annotating the review entities and their attributes in Web pages. The proposed ARROW framework for automatically annotating review pages by adding RDFa annotations to a Web page is composed of four stages: hotspot identification, subjectivity analysis, information extraction, and page annotation.

After normalizing the data, i.e., converting the HTML documents to DOM trees, we continue with identifying the potential reviews or *hotspots* of the page. Usually, reviews are characterized by blocks of text. These blocks are less often found in page headers, navigation elements, footers, etc. Text blocks are usually structured by small amounts of HTML elements, such as `h1` and `div`. Hence, for identifying reviews, we aim to find the elements that contain a lot of textual content. For this, we calculate a text-to-content ratio, the *TTCR*, which can be denoted as

$$TTCR = \frac{L_{text}}{L_{DOM}} , \quad (1)$$

where the number of characters in text is denoted by L_{text} and the total number of characters within the DOM tree is represented by L_{DOM}. HTML elements with a high text-to-content-ratio are labeled as hotspots.

After hotspot detection, we need to verify the hotspots, as they might contain reviews. A review can be defined as a subjective view on a certain topic, as opposed to an objective view which describes only facts about a topic. In order to be able to analyze the hotspots, we use an improved version of the LightWeight subjectivity Detection mechanism (LWD) as proposed by [1], which now also takes into account the length of the review. More precisely, hotspots where a certain number of sentences contain a minimum number of subjectivity words per sentence are considered to represent reviews.

For review attribute extraction we employ several methods. Authors are identified by means of a Named Entity Recognizer (NER), whereas dates and ratings are recognized by means of regular expression patterns. Products are filtered from titles, as it is often hard to identify the product in the review content due to the frequent mentioning of related products.

Finally, after reviews and attributes have been identified in the Web pages, the framework annotates pages using Google's RDFa vocabulary designed by Google for its Rich Snippets. Annotating involves tagging the identified key elements of the review.

4. ARROW EVALUATION

We have implemented the ARROW framework as a Web application[1]. The approach is evaluated on data from various review Web sites[2]. We evaluate the framework on review identification and attribute identification. On average, review annotation is a subsecond process for each Web page.

To assess the review recognition performance, we test the tool on a selection of 100 English review Web pages and 100 non-review English Web pages. When comparing manually annotated reviews with ARROW's annotations, we obtain good results on precision and specificity, yet varying results on accuracy and recall. The results also show us that the framework works better on some Web sites than on others, caused by type of content, specific Web site structures, etc. When performing a similar experiment in order to assess the performance of review attribute identification, we can conclude that our framework does a good job on finding the item reviewed, date, and rating, but performs poorly on detecting the authors. This can be explained by the ambiguity of the names used on the Internet, as many people use nicknames on the Internet rather than their real (full) names. This makes the automatic identification of people difficult.

[1]Available at `http://www.arrow-project.com/`.

[2]Data extracted from `http://www.tripadvisor.com`, `http://www.epinions.com`, `http://www.imdb.com`, `http://www.yelp.com`, and `http://www.cnn.com`.

5. CONCLUSIONS

Using Google Rich Snippets for semantic annotation allows for a more appealing presentation by emphasizing some specific concept properties. Unfortunately, there are not yet many Web sites that support this vocabulary. In order to allow existing Web sites to make use of Google Rich Snippets, we have proposed the ARROW framework in this paper, which aims to automatically identify and annotate reviews on Web pages using Google's vocabulary. We have evaluated an implementation of the framework, which yields good results on precision and specificity, yet varying results on accuracy and recall.

As future work, we suggest to extend our framework to cover other elements from the Google Rich Snippets vocabulary, e.g., recipes, videos, and organizations. Also, one could take into consideration that many reviews lack an explicit rating, e.g., a grade or a number of stars. As Google Rich Snippets accepts a rating based on a scale of 1 to 5, it would be useful to investigate ways of calculating ratings based on review texts using, for example, sentiment analysis methods.

6. REFERENCES

[1] L. Barbosa, R. Kumar, B. Pang, and A. Tomkins. For a Few Dollars Less: Identifying Review Pages Sans Human Labels. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2009)*, pages 494–502. ACL, 2009.

[2] V. Crescenzi, G. Mecca, and P. Merialdo. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. In *27th International Conference on Very Large Data Bases (VLDB 2001)*, pages 109–118. Morgan Kaufmann Publishers Inc., 2001.

[3] R. Khare and T. Çelik. Microformats: A Pragmatic Path to the Semantic Web. In *15th International World Wide Web Conference (WWW 2006)*, pages 865–866. ACM, 2006.

[4] J. van der Meer, F. Boon, F. Hogenboom, F. Frasincar, and U. Kaymak. A Framework for Automatic Annotation of Web Pages Using the Google Rich Snippets Vocabulary. In *26th ACM Symposium on Applied Computing (SAC 2011)*, pages 763–770. ACM, 2011.

[5] J. Wang and F. H. Lochovsky. Data Extraction and Label Assignment for Web Databases. In *12th International World Wide Web Conference (WWW 2003)*, pages 187–196. ACM, 2003.