Automatic Identification of Multipage News: A Machine Learning Approach

Pashutan Modaresi

Heinrich-Heine-University of Düsseldorf
Institute of Computer Science, Düsseldorf, Germany
modaresi@cs.uni-duesseldorf.de

Online news contain valuable information that can be utilized for private or commercial purposes. In the commercial context, online media monitoring services provide other companies or individuals with their required information in a systematic manner. This is accomplished by crawling plenty of news websites. Numerous news websites follow the strategy of pagination to split the stories into multiple pages. Given that, to identify multipage stories, manual rules have to be defined. On the other hand, the dynamic nature of the HTML pages requires a tremendous amount of effort in maintaining these rules. With this in mind, in this work we propose an automatic approach to identify multipage news stories.

We collected a list of web-pages in which the news were splitted in multiple pages and manually annotated them. To each link on the page a label has been assigned. That is, a link either points to the next page of the news or not. As the number of links which do not point to the next pages significantly dominates the number of link pointing to the next page of a news, the data set is highly imbalanced. Moreover, in order to design a language independent algorithm, news pages originating from different countries have been considered.

For each link, the *class* and *id* attributes of the corresponding anchor element, together with the text content of the anchor have been concatenated and fed into a Naive Bayes classifier. The same set of features extracted from the parent elements of an underlying link has been fed into another Naive Bayes classifier. Moreover, the relative position of a link on the news page (calculated by means of a heuristic) has been used to train a regression model. Additionally, some other features such as the structure of the *href* attribute of an anchor or the length of its text content have been integrated. Intentionally, the similarity between the content of the base page and the one of the target page has be ignored, as the calculation of this feature requires network availability that is not always given.

By cause of various learning algorithms being used, the final binary decision has to be performed by combining the results of the single constructed models. For this we use a *stacking* technique where we train a learning algorithm to combine the predictions of the constructed models.

Our first experimental results have revealed very high precision and recall values (≥ 0.9) for both labels under analysis.

Copyright © 2015 by the paper's authors. Copying permitted only for private and academic purposes. In: R. Bergmann, S. Görg, G. Müller (Eds.): Proceedings of the LWA 2015 Workshops: KDML, FGWM, IR, and FGDB. Trier, Germany, 7.-9. October 2015, published at http://ceur-ws.org