Acquiring Metadata from the Web

Milan LUČANSKÝ*

Slovak University of Technology
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
lucansky06@student.fiit.stuba.sk

Our work focuses on mining relevant information from the web pages. Unlike plain text documents, web pages contain another source of potentially relevant information – easily processable mark-up. We propose a method to keyword extraction that enhances Automatic Term Recognition (ATR) algorithms intended for processing plain text documents with analysis of HTML tags present in web documents.

The keyword extraction using an ATR algorithm is domain specific, because with different collections it yields different results. Different ATR algorithms use different measures, which are based either on statistical or probabilistic approach while some algorithms combine both approaches [2]. The extraction of keywords from web pages is even more specific, because they typically cover topics from various domains and a variable number of pages relate to every topic. Such diversity of topics usually reflects into extraction of less descriptive keywords. An approach to keyword extraction from web pages could possibly benefit from other sources of information that the Web offers. The challenge is to consider web mark-up and to make use of emergent semantics that HTML tags represent [1].

Advantages of ATR algorithms, such as the ability to extract most relevant single and multi-word terms or processing plain text, seem to be appropriate for textual content in web environment. But none of them (to our best knowledge) considers the structure of document as potential source of additional information to find the best candidates for keywords. In our approach we combine the ATR algorithms with the processing of HTML tags.

We aim to enhance a way how a final weight of the candidate term is computed. We introduce a *TagRel* coefficient that modifies weight of a term obtained by an ATR algorithm according to the relevance of HTML tag enclosing the term. Our method for keyword extraction consists of the following steps:

1. Web structure preprocessing.
2. Term extraction.
3. Keyword selection.

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering

Spring 2011 PeWe Workshop, April 1, 2011, pp. 13–14.

In the first step we analyze the link structure of examined web pages to obtain information about other pages present "outside" the pages. When examining a particular page, we focus on the anchors of links pointing to that page from the rest of pages in order to extract terms describing the page. We can either crawl pages (in case of closed corpus such as a website) or leverage already existing indices of web search engines (in the case of the open web, e.g., by using '*link:*' operator provided by Google). From the crawled page we also obtain HTML mark-up which emphasizes some words. The most interesting words are visually distinguished from the rest of textual content, such words are enclosed in tags: <*em*>, <*b*>, <*i*>, <*strong*>, <*h1-6*>, <*li*>, etc. Widely used and popular in styling web page content are Cascade Style Sheets (CSS), therefore we analyze the linked .css file, in order to obtain additional information from examined web page. Another possible option is to analyze targeted advertisements placed on some web pages.

Having web structure analyzed, in the second step we extract terms (i.e., candidate keywords) and compute weights reflecting their significance for a document. We modify the weight obtained by an ATR algorithm by *TagRel* coefficient bound to a tag enclosing the term:

$$w_i' = w_i \times TagRel_T \tag{1}$$

where w_i' is improved weight of a term i, w_i is weight of a term i obtained by a ATR algorithm and $TagRel_T$ represents relevance of a tag T that encloses term i.

TagRel varies among different HTML tags. We consider tag importance as an indicator of how much a tag is important for a page as some HTML tags are more likely to contain keywords (terms) than others, therefore their value assigned to *TagRel* will be greater. We have an ambition to parameterize value of *TagRel* according to factors by which an HTML tag or CSS element occurs on web page. The value of *TagRel* will be dependent on the count of occurrences of an HTML tag or the length of phrase emphasized by a tag/css element.

Evaluation of our method will be performed on set of web pages from different domains.

Acknowledgement. This work was partially supported by the Scientific Grant Agency of Slovak Republic, grant No. VG1/0675/11.

References

[1] Hodgson, J.: Do HTML Tags Flag Semantic Content? *IEEE Internet Computing*, Vol. 5, No.1, 20-25, 2001.

[2] Knoth, P., Schmidt, M., Smrž, P., Zdráhal Z.: Towards a Framework for Comparing Automatic Term Recognition Methods, In: *Znalosti 2009,* pp. 83-94, 2009.