Acquiring Web Site Metadata by Heterogeneous Information Sources Processing

Milan LUČANSKÝ*

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
lucansky.milan@gmail.com

We live in world where the amount of freely accessible data increases faster than ever before. World Wide Web is almost unlimited source of knowledge and information and every year the number of available web sites increases in millions. That introduces the demand for automatic processing of vast collection of web documents. We need to assign descriptive metadata to web pages to facilitate further processing and it turns out that keywords are suitable representation of web content. Nowadays, keywords form a basis for semantic representations as they are utilized in the field of ontology engineering [2]. The most of popular search engines are based on keyword search paradigm and keywords are even used in user modeling for adaptive web-based systems to represent the context [1]. Social services, such as Delicious[1] utilize keywords too.

We need an automatic approach to keywords acquisition. In offline document collections there are various approaches to automatic term recognition (ATR). ATR algorithms use statistical and probabilistic features to get relevant keywords and are widely utilized for plain text document (with no internal structure) processing. If used on web documents, they could possibly benefit from hidden semantic of HTML elements used to format and style sheets to visualise text content. Our current research aims at cascade style sheets (CSS) as additional source for identifying potential keywords. The idea of utilization of CSS in co-operation with ATR algorithms is quite new and unexplored. Therefore we see a possibility to combine semantic potential of HTML tags and CSS with ATR algorithms in order to yield better results than using them separately.

We introduce a *TagRel*, *LinkRel* and *CssRel* coefficients that modify weight of a term obtained by an ATR algorithm. Plain text content is passed to ATR algorithm, which extract weighted keywords. From the web page we extract keywords formatted by selected CSS attributes, compute the *CssRel* coefficient and improve ATR

* Supervisor: Marián Šimko, Institute of Informatics and Software Engineering
[1] http://www.delicious.com/

Spring 2012 PeWe Workshop, April 1, 2012, pp. 71–72

keywords weights. For the anchor texts pointing to examined web page we compute *LinkRel* and improve ATR keywords. Finally we acquire text content from selected HTML elements, compute *TagRel* and improve ATR keywords. The equation (1) introduce scheme for computing final weight of extracted keywords.

$$w_t = w_t' + w_t'' \tag{1}$$

where w_t is final weight of a term *t*, w_t' is weight of a term *t* obtained by a ATR algorithm and w_t'' represents weight of term *t* as combination of *TagRel*, *LinkRel* and *CssRel* coefficients according to weighting scheme.

While we use three different coefficients for assigning weight to potential keywords we need a scheme for combining them to the single measure (w_t''). A possible option is to use a weighting scheme. We assign to each type of *Rel* coefficient a multiplication number denoting probability of containing relevant keywords. The estimation of multiplication numbers that should denote the probability of containing relevant keywords is part of the research.

$$w_t'' = \alpha \cdot Link\,\mathrm{Re}\,l + \beta \cdot Tag\,\mathrm{Re}\,l + \gamma \cdot Css\,\mathrm{Re}\,l \tag{2}$$

where w_t'' is combination of *TagRel*, *LinkRel* and *CssRel* coefficients for term *t*, α is the multiplication number for *LinkRel*, β is a multiplication number for *TagRel* and γ is a multiplication number for *CssRel*.

Using equation (1) we produce new order of extracted keywords, where the most relevant should have the highest weights. In order to evaluate the proposed approach, we are currently conducting an extensive experiment on a set of randomly chosen pages from the World Wild Web. So far we performed synthetic experiment on small set of randomly chosen web pages, in order to process visual information represented by style sheets formatting. The web pages use different style sheets formatting. We extracted all emphasized words and short terms from main textual content and tried to state either the term is relevant to the contents of web page or not. In average 38 % of extracted terms were relevant to the topic of article. Actual results seem very encouraging. In a more extensive experiment we need to examine the method on different types of web pages and to compare results with cotemporary approaches (e.g., freely available web services for term extraction as tagthe.net[2], OpenCalais[3]).

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

References

[1] Barla, M., Bieliková, M. Ordinary Web Pages as a Source for Metadata Acquisition for Open Corpus User Modeling. In White,B., Isaías, P., Andone, D., (Eds.): WWW/Internet 2010, IADIS Press, pp. 227-233

[2] Cimiano, P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Springer-Verlag, 2006, pp 23-24.

[2] http://www.tagthe.net/

[3] http://www.opencalais.com/