Acquiring Metadata about Web Content Based on Microblog Analysis

Tomáš UHERČÍK *

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 3, 842 16 Bratislava, Slovakia
uhercik07@student.fiit.stuba.sk

The amount of information on the Web is so huge, that searching can be done only by machines. However, information presented on the Web is intended for humans and is understandable only by humans. The Semantic Web is vision, where this problem is solved by the layer of machine-processable metadata. These metadata are not available as often as we would like. The challenge is to obtain them automatically.

Socially-oriented data are those, which are created by the activity of users. There is a lot of useful metadata within that data. Web applications for social networks allow a user to share a lot of information with others. Data created by their activity are very valuable source of indirectly originated metadata.

We decided to use the microblog *Twitter* as source of metadata. We selected the URL as entity about which are the metadata acquired, because it can be unambiguously identified in the tweets' text.

We proposed a method for keyword extraction utilizing *Twitter* posts. Its flow is illustrated in the Figure 1.

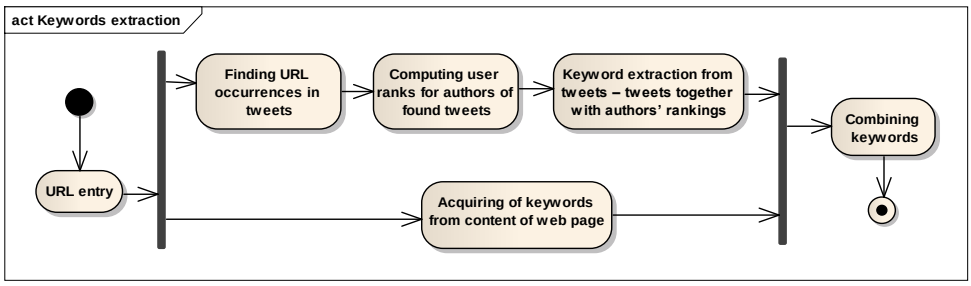

Figure 1. Activity diagram showing the process of the proposed method.

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

Spring 2012 PeWe Workshop, April 1, 2012, pp. 83–84.

In addition to ordinary extraction methods we consider the different relevance of particular tweets depending on an author who published them. It is very effective to use this information as input of extraction method. We proposed the ranking formula, based on Tunkrank [1] with dependency on the frequency of user's tweets publishing as follows:

$$URank(X) = \sum_{Y=followers(X)} \frac{1+\frac{p}{log(T)}.URank(Y)}{|followers(Y)|} \quad (1)$$

where *URank* is the user(author) ranking, *X,Y* are users, *followers(X)* is the set of users following *X*, *p* is convergence constant and *T* is median of time gaps between publishing individual tweets.

For extraction of relevant keywords we used TextRank algorithm [2], but we could use any extraction algorithm, which would give us keywords with their relevancies. Final *Twitter* (microblog) relevance *MRank* of keywords we obtain as follows:

$$MRank(t) = \max(URank(t)) * TRank(t) \quad (2)$$

where the *max(URank(t))* is maximum of all user ranks of all users, who are authors of tweets, which contains extracted keyword and *TRank* is the textual relevance of keyword.

For evaluation, we obtained more than 50 GB dataset from *Twitter* using *Twitter streaming API* during 10 days. We evaluated the results of our method for the set of 10 recent URL from Twitter. We obtained average precision 86 %. We also measured to what extent our method enriched basic set of keywords extracted from resource content only. The Twitter keyword is important, when is relevant and we cannot find it within keywords extracted directly from content of URL. 46% of extracted keyword matched this condition. We consider this enrichment very reasonable and metadata coming from *Twitter* to be very valuable.

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

References

[1] Tunkelang D.: A Twitter Analog to PageRank, 2009. Available at: http://thenoisychannel.com/2009/01/13/ a-twitter-analog-to-pagerank/ [cit. 2011-11-6]

[2] Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: *Proc. Of Conf. on Empirical Methods in Natural Language Processing*, 2004, ACL, pp. 404-411.