Zero-cost Labelling with Web Feeds for Weblog Data Extraction

George Gkotsis*, Karen Stepanyan, Alexandra Cristea, and Mike Joy
Dept. of Computer Science, University of Warwick, UK

Volume
[2008] 900 posts/day
[2009] 1334 English blogs in one month
(2011) 25% of Internet users in Britain have a blog

Diversity
• 1634 WordPress themes
• 469 different platforms

Blogosphere
Web data extraction is a genuinely hard problem. The blogosphere, which constitutes a constituent part of the Web, remains bound to the limitations of modern data extraction. In general, data extraction is facing the trade-off between automation and accuracy/granularity. The proposed model overcomes the above limitations by exploiting an inherent characteristic of weblogs: the Web Feed, commonly provided as RSS.

Proposed Methodology

Filter generation

Step 1: Feed Processing
1. Reading of post properties
2. Fetching of learning posts

Step 2: Filter generation

Filter repository

Step 3: Wrapper execution

The last step transforms the filters into rules, in order to calculate the scores and select a rule for each of the desired properties. Essentially, a rule is the result of the transposition of a filter. This transposition can result in maximum three rules. Hence, a rule is described by its type (one of the three different attribute types of the filters), a value (the value of the corresponding filter’s attribute) and a score, which is used to measure its expected accuracy. The need to calculate the score of each rule is justified by the inherent “noise” of the filters.

Evaluation

The rules extracted show that the rule types vary for different weblog properties.

* gkotsis@gmail.com
gkotsis@twitter

This work was conducted as part of the BlogForever project funded by the European Commission Framework Programme 7 (FP7)