Automated Recognition of Writing Style in Blogs

Martin VIRIK*

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

In the current web, blogs represent a genre that stands between static pages and live forums. They have became a tool for ordinary users to share information, ideas or even emotions, creating a heterogeneous mass of user generated content filled with unique information related to individual as well as society-wide issues. Since the blog articles are typically weakly structured, the extraction of valuable information is becoming increasingly difficult to handle.

Even though there are no restrictions or limits to content or form in blogs, users seem to spontaneously create writing styles and genres reflecting their intention and current emotional state. Modern search services are aware of these genres and consider them as highly valuable for several tasks. For example, for filtering articles with low information value or recognizing sentiment about a specified object.

In our research we follow a distribution between informative and affective articles [1] and for affective articles we suggest two dichotomies of further differentiation: reflective vs. narrative and emotional vs. rational. By their combination, we receive four categories, which we will use for our classification (direct reaction on an event, rational reflective article, direct reaction on writer's day or a story from past, i.e. vacation, a tale from childhood). We have gathered a dataset for initial experiments of about 16 thousand blog posts and managed to manually classify a subset in a user experiment including several participants.

In our work we focus on linguistic characteristics of blog articles. We propose a novel method for Slovak blogs classification that considers not only word usage and lexical and morphological attributes (typical for state-of-the-art approaches [2]), but also more complex features such as sentence syntax or text structure obtained during a pre-processing step. We have improved the morphological tagging by considering word's position in sentence enabling syntactic analysis, as well as capturing the usage of each word class. We have built a lightweight syntactic analyzer, which transforms morphologically tagged text into a structure based on an object model depicted in

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

Spring 2011 PeWe Workshop, April 1, 2011, pp. 71–72.

Figure 1. A result of this transformation is the localization of predicate candidates, which are the verbs that we have used to identify the sentences in the previous step. In our feature set we investigate their dominant tense, person and number categories along with the intensity of this dominance.

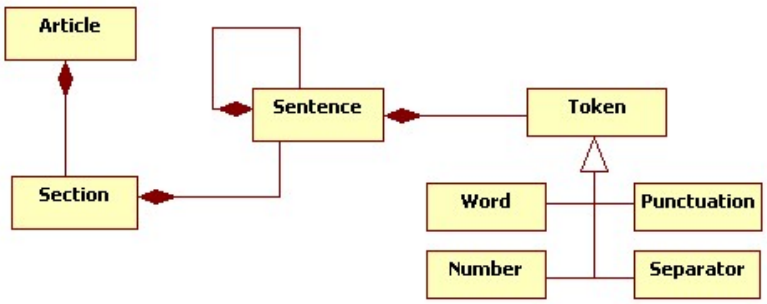

Figure 1. Class diagram of article structure model.

Besides the simple structural features (i.e. number of sections, average number of sentences per section) we measure also the modified standard deviation of section length. We have modified the standard deviation formula by counting the proportion of difference and average, so now it reflects the variation of section length and the consistency of the whole text.

To evaluate our method, we have suggested a compound classifier based on three binary classifiers for each pair of classes. We have conducted an initial experiment, in which each article was represented by a 28-dimensional feature vector and for each classifier we have integrated and configured Naïve Bayes (as baseline classifier), Suport Vector Machine and k-Nearest Neighbours classifiers from Weka tool. By applying 10-fold cross-validation evaluation method we have gathered first result with precision 68–81%.

In the current phase of our project we are gathering a larger training set of manually classified articles and improving lightweight syntactic parsing methods and methods for feature selection. We are experimenting with classification algorithms and classifier comities in order to boost the accuracy of classification.

Acknowledgement. This work was partially supported by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 345-032STU-4/2010.

References

[1] Ni, X., Xue, G., Ling, X., Yu, Y., Yang, Q.: Exploring in the Weblog Space by Detecting Informative and Affective Articles. In *Proc. of 16th Int. Conf. on World Wide Web*, ACM, pp., 281–290, 2007.

[2] Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. *Communications of the ACM*, Vol. 52, No. 2, 119–123, 2009.