Automated Recognition of Author's 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 past decade it has become much easier to create web content even for users with no experience with web technologies. Weblogs are the most typical and the most growing example of this trend. Thousands of bloggers use this hybrid genre to express their ideas, opinions and emotions, making blogs a rich space of topics and writing styles. In proportion to increasing number of blogs, the number of efforts to improve blog-search and recommendation algorithms has also grown. New requirements are aware of blog articles text quality and consider individual writing style an important blog characteristic.

In our research we focus on linguistic characteristics of blog articles in order to recognize and classify writing style of articles, blogs or even authors. We study the grammar and morphology of selected language and possibilities of computational linguistics to extract the features of document model necessary for further classification. In the first phase of our research we have been studying basic text mining and classification methods [1] and works related to the analysis of blog articles linguistic quality. Apart from user profiling, such as gender or personality profiling [2], a great effort has been on differentiating between informative and affective articles [3]. This and other genre based research has proven a large overreach of affective blogs, especially diaries. Methods analyzing reading difficulty are much related to the weblog classification [4]. We discovered a space for building models for multiple factors such as measures of syntactic complexity or prior knowledge of the reader.

We plan to build on system architecture for an advanced text mining system with background knowledge base as described by Feldman & Sanger [1] (see Figure 1). In our research we aim to accomplish preprocessing tasks and to create the processed article collection. This collection will be used by text mining algorithms, which discover patterns and trends and respond to user requests by considering his background knowledge and preferences.

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

Spring 2010 PeWe Workshop, April 18, 2010, pp. 37–38.

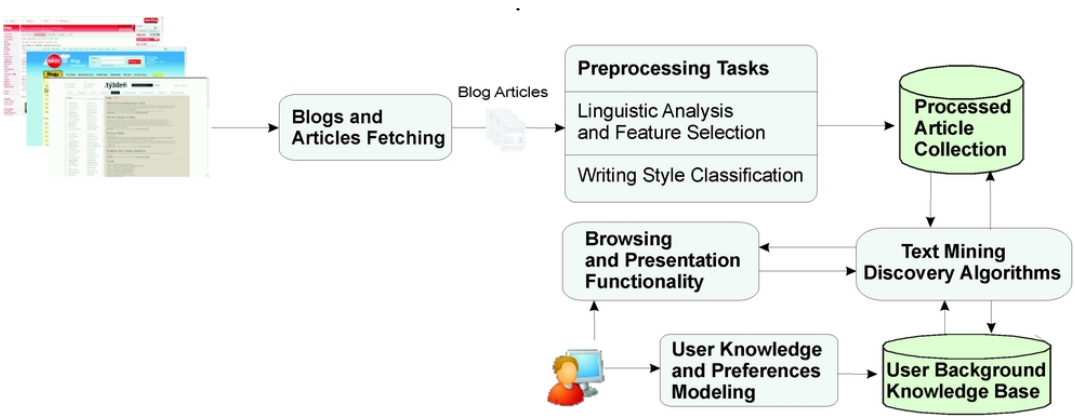

Figure 1. Proposal of system architecture for blog writing style classification (based on [1]).

We have considered many different types of features as potential markers of writing style, including lexical and syntactic complexity-based features or features mapping vocabulary difficulty level.

Average word length and syllables count are standard features measured in most readability indexes, such as ARI, Gunning-Fog Index or Flesch Reading Ease. Algorithms behind these indexes are relatively fast and easy to optimize to any alphabetic language. Together with grammatical and orthographic error count in unedited articles, these features could create a good picture of blogs lexical quality.

Capturing syntactic complexity of informal blog articles could bring more sophisticated view on text structure. Using advanced algorithms, such as Linked Grammar, can create space for discovering syntactic patterns, which could be used to characterize reading difficulty of writing style.

We plan to evaluate the results by applying our method on the collection of articles from live blogs and gathering implicit and explicit feedback from users.

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] Feldman R., Sanger J. Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, New York. 2007

[2] Argamon S., Koppel M., Pennebaker J.W. and Schler J. Automatically profiling the author of an anonymous text, Communications of the ACM, v.52, n.2, p.119-123, February 2009

[3] Ni X., Xue G., Ling X., Yu Y. and Yang Q. Exploring in the Weblog Space by Detecting Informative and Affective Articles. Comparative and General Pharmacology, p. 281-290. 2007

[4] Miltsakaki E., Troutt A. Real-Time Web Text Classification and Analysis of Reading Difficulty. Computational Linguistics, p. 89-97. June 2008