Characterizing Weblog Corpora

Fernando Perez-Tellez[1], David Pinto[2], John Cardiff[1], Paolo Rosso[3]

[1]Social Media Research Group, Institute of Technology Tallaght, Dublin, Ireland
fernandoperez@itnet.ie, John.Cardiff@ittdublin.ie
[2]Benemerita Universidad Autónoma de Puebla, Mexico
dpinto@cs.buap.mx
[3]Natural Language Engineering Lab. – EliRF, Dept. Sistemas Informáticos y
Computación, Universidad Politécnica Valencia, Spain
prosso@dsic.upv.es

1 Introduction

In order to exploit the huge volume of information being published in the blogosphere, it is essential to provide techniques such as clustering, which can automatically analyze and classify their contents. However these typically can produce better results when dealing with wide domain full-text documents. In most cases however, blogs can be considered to be "short texts", i.e., they are not extensive documents and exhibit undesirable characteristics from a clustering perspective such as low frequency terms, short vocabulary size and vocabulary overlapping of some domains. Furthermore, their characteristics vary widely depending on the specific interests of the writer, their linguistic style, and the volume of texts that they produce.

In this work, we present a set of evaluation features by which we can establish the relative hardness of the clustering task, i.e., how easy or difficult it will be to accurately cluster the blog datasets. These are the shortness, domain broadness, class imbalance, stylometry, and structure. We report results obtained on corpora extracted from two popular blogging sites, Boing Boing ("B-B") and Slashdot[1]. The results are contrasted with characterizations of a number of other corpora, consisting of newspaper articles and academic papers. We can use the results to provide knowledge of the most appropriate methodology for clustering.

2 Corpora Evaluation Measures

As established in [2], it is important for the evaluation measures to be domain-independent. We focus on the following corpus characteristics: *Domain Broadness* detects if two different categories should be tagged as "wide" or "narrow". *Class Imbalance* indicates how evenly the documents are distributed across the categories. *Stylometry* helps to distinguish between writings styles of authors. *Shortness* calculates features derived from the length of a text, such as the maximum term frequency per document. *Structure measures* validate the similarity and dissimilarity

[1] Boing Boing http://boingboing.net; Slashdot http://slashdot.org. A preprocessed version of each dataset is available at http://www.dsic.upv.es/grupos/nle/downloads.html (June 2009)

of the categories of the gold standard by providing a single value which represents the structure of the document collection. By determining the degree of these measures of corpora we can test clustering methods in order to determine the complexity of classifying text collections of this type.

3 Experiments and Results

In Table 1, we compare the results of the evaluation of the blog corpora with four other short text corpora, consisting of scientific abstracts and newsfeeds [2].

Table 1. Features of the corpora.

Corpora	Text type	Shortness	Class imbalance ranking	Stylometry	Structural ranking	Broadness measures
Cicling-2002	Scientific	Very short	4	Specific	4	Narrow
WSI-SemEval	Scientific	Very short	1	Specific	3	Narrow
R8-Training	News	Short	3	General	2	Wide
R8-Test	News	Short	2	General	1	Wide
B-B	Blog	Very short	4	General	4	Narrow
Slashdot	Blog	Very short	4	General	3	Narrow

In relation to class imbalance, both blog corpora appear well-balanced and so this will have no impact on the clustering blog process. The stylometry measure indicates a general language writing style for blog corpora as the documents were written by many different people. The structural measures confirm the best gold standard structure is for Slashdot. We validate the similarity and dissimilarity of the suggested groups or categories of the gold standard providing a single value which represents the structure of the document collection. Slashdot seems to have good structure, it is expected that Slashdot obtain much better clustering results than B-B. In relation to broadness, Slashdot is shown by all measures to be of narrower domain than B-B but both are considered narrow domain.

4 Conclusions and further work

In the corpora we have analyzed, our experiments indicate that the blogs were characterized as short text, with a general writing style and in a narrow domain. Having this knowledge allows us to employ a methodology for clustering which is described in [1] which takes advantage of a new self-enriching technique in order to address the challenges of clustering short texts, particularly in relation to blogs.

Acknowledgements: The work of the first author is supported by the HEA under grant PP06TA12. The work of the fourth author is supported by the CICYT TIN2006-15265-C06 research project.

References

1. Pinto D., Rosso P., Jiménez-Salazar H., UPV-SI: Word Sense Induction using Self-Term Expansion. 4th. Workshop on Semantic Evaluations - SemEval 2007. Association for Computational Linguistics (2007)

2. Pinto D., On Clustering and Evaluation of Narrow Domain Short-Text Corpora, PhD dissertation, Universidad Politécnica de Valencia, Spain (2008)