Defining and Evaluating Blog Characteristics

Fernando Perez-Tellez
Social Media Research Group
Institute of Technology Tallaght
Dublin, Ireland
fernandopt@gmail.com

David Pinto
B. Universidad Autónoma de Puebla
Puebla, Mexico
dpinto@cs.buap.mx

John Cardiff
Social Media Research Group
Institute of Technology Tallaght
Dublin, Ireland
John.Cardiff@ittdublin.ie

Paolo Rosso
NLE Lab – EliRF–DSIC
Universidad Politécnica de Valencia
Valencia, Spain
prosso@dsic.upv.es

***Abstract*— The analysis of weblogs has become a popular area of natural language processing. Due to their specific characteristics, such as shortness, vocabulary size and nature, etc. it can be difficult to achieve good results using automated clustering techniques. In particular, their nature can vary considerably, both in length and in breadth of topic. Without a priori knowledge of the nature of a blog it is difficult to achieve accurate clustering results. In this paper, we present a framework for the assessment of a set of corpus features that will provide us with insight into their nature from a number of perspectives including shortness, broadness and class imbalance. This in turn allows us to assess the relative hardness of the clustering task and to identify components that can improve the accuracy of the clustering task. We furthermore present the results of some experiments in which we analyzed the features of two sample blog corpora, and we compared the results with other kinds of short texts.**

Keywords: Blogs, Characterization, Short text.

I. INTRODUCTION

Recent years have seen the World Wide Web being used increasingly as a tool in the process of socialization with services such as blogs and wikis, being developed under the "Web 2.0" umbrella. Blogs have become a particularly important decentralized publishing medium. They are reverse chronological sequences of highly opinionated and personal on-line commentaries which make it possible for a large number of people to share their ideas and spread opinions over the Internet. There is considerable interest in developing computational approaches to the analysis of blogs for harnessing their contents in a variety of spheres including gender and sentimental analysis, marketing, government interest, and education. For instance, [12] presents an approach to find happiness and sadness factors in blogs. In [14] the authors describe an approach to analyze gender preferences by mining the blogosphere.

In order to manage the huge amount of information published in the blogosphere, techniques are required which can analyze and classify automatically the information itself, and provide useful information for their processing by information retrieval systems. Methodologies for document clustering – the assignment of documents to previously unknown categories – have traditionally been employed to this end. However these approaches/techniques typically can produce better results when dealing with wide domain full-text documents because the algorithms can deal more effectively with more discriminative information. Commonly blogs are commonly considered as "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 breadth of topics on which they write), their linguistic style, and the volume of texts that they produce.

In order to improve the quality of the clusters produced, it is necessary to assess those features of the texts which can impact the effectiveness of the clustering procedure. We can use the results to provide knowledge of the most appropriate methodology for clustering [8]. The aim is to reveal features which could predict the success of clustering a given corpus.

We consider a necessity to employ a detailed analysis of features of blogs in order to be more objective. Additionally, we attempt to isolate characteristics which are unique to blogs by comparing our results to corpora of scientific abstracts and short news articles.

In this paper, we present a set of evaluation features which can characterize blogs. These evaluation features may be used to establish the relative hardness of the clustering procedure, in other words, how easy or difficult it will be to accurately cluster the blog datasets. In particular we focus on the shortness, domain broadness, class imbalance and stylometry. 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. In general, we expect to be able to differentiate between texts of scientific paper abstracts and blogs in order to predict the best manner to cluster those corpora and obtain the best results. By determining the degree of broadness, stylometry, class imbalance and shortness of corpora we can test clustering methods in order to determine the complexity of classifying text collections of this type. This will enable us to analyze and propose an appropriate methodology that could improve the obtained accuracy in the clustering task [8].

[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)

The remainder of this paper is organized as follows. In the next section, corpora used in this work are presented. In Section 3 the corpora evaluation measures are introduced. Section 4 describes the evaluation results. Finally in Section 5 conclusions and further work are discussed.

II. CORPORA

In this section we describe the main characteristics of the blog and reference corpora. We have preprocessed all these collections by eliminating stop words and by applying the Porter stemmer [13]. The statistics quoted were obtained after application of this pre-processing.

A. Blog Corpora

Table I presents some properties of the two datasets being characterized. These properties include running words (number of words in the corpus), vocabulary size, number of post categories, number of discussion lines, and the total number of posts. We consider discussion lines to be those used when the posts were manually classified (eg., D1, D2, …, D12 in Figure 1, in the case of the Boing Boing corpus). The posts are comments generated by a specific discussion line.

TABLE I. PROPERTIES OF THE BLOG DATASETS

Corpus property	Boing Boing	Slashdot
Running words	75935	25779
Vocabulary size	15282	6510
Post categories	4	3
Discussion lines	12	8
Posts	1005	8

The gold standard for B-B and Slashdot (i.e., the manually constructed expert classification) was created by Perez [6]. In the construction of the B-B corpus, each post in each discussion line was labeled with its number of discussion line and the number of the post. In Figure 1, we have presented an example of some discussion lines in order to easy understanding of the corpora structure. For example, "D1" indicates discussion line number one, with topic "Toronto's science fiction reading series; launching my little brother on May 1", and category "Book".

- **D1** Toronto's science fiction reading series; launching my little brother on May 1 ### Book
- **D2** Behind TV "military analysts," the Pentagon's hidden hand ### New

…

- **D11** 2001 profile of "Bill Ayers, unrepentant former Weather Underground revolutionary" ### New
- **D12** Nelson Mandela and the ANC are on the US terrorist watchlist and need waivers to enter the country ### Civlib

Figure 1. Example of Boing Boing discussion lines

Figure 2 shows some posts, each of which is identified by the number of the post and the discussion line to which it belongs; for instance, P2D1 means post two (P2) in the discussion line one (D1). Each identifier is followed by the message and the date and time at which it was posted.

B. Reference Corpora

The CICLing-2002 corpus. This corpus is made of 48 documents from the Computational Linguistics domain, which corresponds to the CICLing 2002 conference [2]. Although it is very small, it is useful as a reference corpus because we may manually verified the obtained results. The features of this corpus are shown in Table II.

- **P2D1** April 31 does not appear on my calendar. Maybe it's a Canadian thing? Take a look at this ### jonesey ### April 19 2008 6:36 AM
- **P3D1** Perhaps you mean May 1st, not April 31st? I was looking forward to seeing a bunch of the sf authors at Foresight.. but the TTC strike is going to make that difficult. Take a look at this ### Whiskeyjack ### April 19 2008 6:37 AM
- **P4D1** April 31 is when all the events that were promised on April 1 take place. The book's official launch is on April 28, but that's in the US. I'm not sure what Cory meant here. Take a look at this ### Xopher ### April 19 2008 6:45 AM

Figure 2. Example of Boing Boing posts

TABLE II. FEATURE AVERAGES OF THE CICLING-2002 CORPUS

Feature	Full documents	Abstracts
Number of categories	4	4
Number of abstracts	48	48
Total number of terms	80,109	3,382
Vocabulary size (terms)	7,590	953

The WSI-SemEval collection. This data collection was provided by the organizers of the "Evaluating Word Sense Induction and Discrimination Systems" task of the SemEval 2007 workshop of the Association for Computational Linguistics[3].The dataset consists of 100 ambiguous words (65 verbs and 35 nouns) borrowed from the "English lexical sample" task of the same workshop. The documents come from the Wall Street Journal corpus, and they were manually annotated with OntoNotes senses. In Table III we show the general features of the *WSI-SemEval* data collection.

TABLE III. FEATURE AVERAGES OF THE WSI-SEMEVAL DATA COLLECTION

Feature	Value
Number of sentences	27,132
Minimum number of categories (senses)	1
Maximum number of categories (senses)	11
Average number of categories (senses)	2.87
Total number of terms	1,555,960
Vocabulary size (terms)	27,656

[2] http://www.cicling.org
[3] http://nlp.cs.swarthmore.edu/semeval/tasks/task02/description.shtml

Reuters-215788[4]. This corpus has been extensively used for categorization tests. The most used version of Reuters is distributed as Reuters RCV1 and RCV2. In the experiments we have carried out, we have used the R8 sub-collections of Reuters-21578 since they are a single-label categorized dataset. Since the R8 corpus is used for the categorization task, it is usual to work with a training and a test version of the data. The characteristics of the R8 corpus are given in Table IV.

TABLE IV. FEATURE AVERAGES OF THE R8-REUTERS DATA COLLECTION

Feature	**Training**	**Test**
Number of categories	8	8
Number of documents	5,839	2,319
Total number of terms	416,431	150,430
Vocabulary size (terms)	15,648	9,315

III. CORPORA EVALUATION MEASURES

Evaluation of linguistic resources is a very important area and it needs to be addressed by international forums. For instance, in [1] is described various measures for evaluating corpus similarity and a strategy for evaluating the measures. Commonly in natural language processing competitions we assume that the corpora is of sufficient quality to be used as a standard in the experiments but it does not imply one hundred percent usefulness or applicability of the resource for specific purposes for which it was constructed. In addition we are working with blogs that by their nature possess particular features that can influence the clustering task itself such as style of writing, low frequency terms, short vocabulary size and vocabulary overlapping of some domains. Ad-hoc clustering methods [8] may be used in order to improve the quality of the obtained results. Therefore, we believe that this study would be highly beneficial.

As established in [7], domain-independent evaluation measures are recommended. In particular, we have used those that take into account the following corpus characteristics: domain broadness, class imbalance, stylometry and shortness. By determining the degree of broadness, shortness and class imbalance of corpora, we may draw conclusions about the value of choosing any particular clustering method. This will enable us to analyze the possible methodologies, and select the method likely to produce the best clusters.

In this section we describe and present the measures for each of the corpora evaluation features we have considered.

A. Domain broadness evaluation measures

The goal of establishing domain broadness degree is to evaluate the broadness of a given corpus from a vocabulary-based perspective. For example, if we have different categories like games and politics, we need a measure to know that these two categories have to be tagged as "wide" categories; on the other hand the categories like games and sports have to be evaluated as "narrow" categories. In general, clustering algorithms produce better results when dealing with broad, rather than narrow domains.

In our experiments we have used two approaches, one based on statistical modeling, and the other based on vocabulary dimensionality.

1) Statistical language modeling

The goal of Statistical Language Modeling (SLM) is to build a statistical language model in order to estimate the distribution of words/strings of natural language. The calculated probability distribution over strings *S* of length *n*, also called n-grams, attempts to reflect the relative frequency in which *S* occurs as a sentence. In this way, from a text-based perspective, such a model tries to capture the writing features of a language in order to predict the next word given a sequence of them.

In our particular case, we have considered that every hand-tagged category of a given corpus to be clustered has a language model. Therefore, if this model is very similar to the other models which were calculated for the other categories, then we could affirm that the corpus is narrow domain. The degree of broadness may be approximated by evaluating this proposed supervised approach over several corpora. Our proposal approaches in an unsupervised way the problem of determining the domain broadness of a given corpus. In fact, we calculate language models for *v* partitions of the corpus without any knowledge about the expert document categorization (gold standard).

The supervised domain broadness evaluation measure is described as: Given a corpus *D* with a gold standard made up of *k* classes $C^*=\{C_1^*, C_2^*, \ldots, C_k^*\}$. We obtain the language model of all the classes except C_i^* $(\vec{C}_i^*)$ and, thereafter, we compute the perplexity [5] of the obtained language model with respect to the model of C_i^*. That is, we use the class C_i^* as a test corpus and the remaining ones as a training corpus in a leave one out process. Formally, the Supervised Language Modeling Based (SLMB) approach for determining the domain broadness degree of the corpus *D* is shown in Eq (1).

$$SLMB(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(Perplexity\left(C_i^* \mid C_i^*\right) - \mu\left(Perplexity\left(C^*\right)\right)\right)^2} \quad (1)$$

Where, $\mu\left(Perplexity\left(C^*\right)\right) = \dfrac{\sum_{i=1}^{k} Perplexity\left(C_i^* \mid C_i^*\right)}{k}$ (2)

The unsupervised Language Modeling Based (ULMB) approach for assessing the domain broadness of a text corpus is calculated as follows. Given a corpus *D* split into

[4] http://www.daviddlewis.com/resources/testcollections/reuters21578/

subsets C_i of l documents, we calculate the perplexity [5] of the language model of C_i with respect to the model of a training corpus composed by all the documents not contained in C_i ($\bar{C}_i^*$).

Formally, given $\bar{C}_i^* \cup C_i^* = D$ such as $\bar{C}_i^* \cap C_i^* = \phi$ and $k = Integer\left(|D|/|C_i|\right)$ with $|C_i| \approx 1$, the unsupervised broadness degree of a text corpus D may be obtained as shown in Eq (3)

$$ULMB(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(Perplexity\left(C_i \mid C_i\right) - \mu\left(Perplexity(C)\right)\right)^2} \quad (3)$$

Where, $\mu\left(Perplexity(C)\right) = \dfrac{\sum_{i=1}^{k} Perplexity\left(C_i \mid C_i\right)}{k}$. (4)

2) Vocabulary dimensionality

This measure of domain broadness assumes that corpora subsets which belong to a narrow domain share the maximum number of vocabulary terms compared with those subsets which do not. In case of a wide domain corpus, it is expected (at least with short texts) that the standard deviation of vocabularies obtained from subsets of this corpus (with respect to the full corpus vocabulary) is greater than the one of a narrow domain corpus. The supervised approach is defined as: given a corpus D with a gold standard made up of k classes $C^*=\{C_1^*, C_2^*, ..., C_k^*\}$. If $|V(D)|$ is the cardinality of the complete document set vocabulary and $|V(C_i^*)|$ the vocabulary size of the class C_i^*, the Supervised Vocabulary Based (SVB) measure for the domain broadness of D may be written as shown in Eq(5).

$$SVB(C) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\frac{|V(C_i^*)| - |V(D)|}{D}\right)^2}. \quad (5)$$

The unsupervised version of the Vocabulary-Based (UVB) domain broadness evaluation measure may be also proposed, this approach would be useful when the gold standard is not available. Since the classes are unknown, we could then use each document instead of the corpus classes. Formally, given a corpus made up of n documents $D=\{d_1,d_2,...,d_n\}$, if $|V(D)|$ is the cardinality of its vocabulary and $|V(d_i)|$ the vocabulary size of the document d_i then the unsupervised broadness evaluation measure of D (based on the vocabulary dimensionality) may be written as shown in Eq(6)

$$UVB(C) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|V(d_i)| - |V(D)|}{D}\right)^2}. \quad (6)$$

B. Stylometry-based evaluation measure

Stylometry studies the linguistic style of a human writer. In [1] is described an approach that uses the occurrence patterns of terms in different collections. One of the practical applications is to determine the authorship of documents. Our aim however is to distinguish between blog writings and other kind of short text.

Our approach is based on Zipf's law [11], which takes into account the term frequency distribution in a document. We have restricted our analysis to determine whether or not a corpus is written by a group of persons with the same linguistic style. It expected that the blog corpora will follow a similar pattern in this measure.

Formally, given a corpus D with vocabulary $V(D)$, we may calculate the probability of each term t_i in $V(D)$ as shown in Eq(7) and the expected Zipfian distribution of terms as shown in Eq(8). We used the classic version of the Zipf's law and, therefore, s was set to 1.

$$P(t_i) = \frac{tf(t_i, D)}{\sum_{t_i \in V(D)} tf(t_i, D)}. \quad (7)$$

$$Q(t_i) = \frac{1/i^s}{\sum_{r=1}^{|V(D)|} 1/r^s}. \quad (8)$$

The unsupervised Stylometric Evaluation Measure (SEM) of D is obtained by calculating the asymmetrical Kullback-Leibler distance of the term frequency distribution of D with respect to its Zipfian distribution, as shown in Eq(9). A general interpretation is high value refers very specific style, whereas a low value would indicate a general language writing style.

$$SEM(D) = \sum_{t_i \in V(D)} P(t_i) \log \frac{P(t_i)}{Q(t_i)}. \quad (9)$$

C. Shortness-based evaluation measures

This evaluation measure calculates features derived from the length of a text, such as the maximum term frequency per document. The term frequency, for instance, is crucial for the major of similarity measures. When dealing with very short texts, we expect that the frequency of their vocabulary terms will be very low. Therefore, the clustering algorithms will have problems for detecting the correct classification, since the similarity matrix will have very low values.

Given a corpus made up of n documents $D = \{d_1, d_2, ..., d_n\}$, we present an unsupervised text length-based evaluation measures which take into account the level of shortness [10]. We directly calculated the arithmetic mean of Document Lengths (DL) shown in Eq(10). As we can see the higher the value is, the longer the text is.

$$DL(D) = \frac{1}{n}\sum_{i=1}^{n} |d_i|. \quad (10)$$

D. Class imbalance degree assessment measure

The assignment of documents to categories allows us to

identify those corpora with almost the same number of documents in each category as balanced; where the number varies considerably it is described as unbalanced. The class imbalance degree is an important feature that must be considered when corpora are categorized, since according to the imbalance degree there could exist different levels of difficulty.

For our purposes, we use a new supervised class imbalance evaluation formula. First, we assume that given a corpus *D* to be categorized with a pre-defined gold standard made up *k* classes ($C^*=\{C_1^*, C_2^*, ..., C_k\}$), the Expected Number of Documents per Class (ENDC) will be:

$$ENDC(D) = \frac{|D|}{k}. \quad (11)$$

The aim of the proposed measure is to determine the Class Imbalance (CI) degree of a supervised corpus which has a gold standard. Thus, the supervised measure is calculated as the standard deviation of *D* with respect to the expected number of documents per class in the gold standard as shown in Eq(12). Here, the higher the value, the more unbalanced the corpus is and vice versa.

$$CI(D) = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(|C_i^* - ENDC(D)|)^2}. \quad (12)$$

IV. EVALUATION RESULTS

In this section, we present the results of our experiments, which are summarized in Table V, the preliminary results were presented in a short version paper in [9]. In each case, we present an analysis and discussion of our results. Evaluation of the abstract text and short news corpora has already been performed [7] and, therefore, we are able to establish a comparison between the previously obtained results and the blog corpora.

TABLE V. RESULTS OF FEATURE ANALYSIS.

	Shortness measures	Class Imbalance Measure	Stylometry	Broadness measures			
Corpora	**DL**	**CI**	**SEM**	**SLMB**	**ULMB**	**SVB**	**UVB**
B-B	84.16	0.02	0.07	92.50	68.61	6.85	7.92
Slashdot	3222.62	0.06	0.07	27.24	33.50	1.06	1.48
Cicling-2002	70.46	0.03	0.30	38.92	63.62	1.73	2.70
WSI-SemEval	59.58	0.22	0.44	195.02	130.62	1.80	3.06
R8-Training	66.32	0.17	0.14	603.95	135.87	3.67	4.76
R8-Test	60.05	0.16	0.09	545.69	134.60	3.84	4.89

In Table VI we can see the features of the reference corpora, for example the *Cicling-2002* corpus is evaluated as being a very short text corpus, with very well balanced classes. *Cicling-2002* is considered as being a narrow domain corpus and regarding with the stylometry category, this text belongs to scientific text. *WSI-SemEval* is another very short text corpus; in this case it is the most imbalanced reference corpus. The *WSI-SemEval* corpus is also included in the stylometry category named 'specific', which implies that this corpus is considered to be a narrow domain corpus. Finally *R8-Training* is considered as short text with general stylometry (as we would expect since this corpus is a news collection), and it is established to be in the category 'wide domain'.

TABLE VI. FEATURES OF THE REFERENCE CORPORA.

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

A. *Document Length Measure*

The Document length (DL) measure reflects the ratio between the number of documents and the size of the documents. The value is smaller for B-B than Slashdot, signifying that Slashdot is composed of a small number of large documents in contrast with other corpora. The obtained values suggest to us that we will obtain better clustering performance with B-B and Slashdot. In the Slashdot corpus, discussion lines were used as categories in its gold standard, which is why it has the highest DL value. When dealing with very short texts, the frequency of their vocabulary is very low and, therefore, the clustering algorithms have the problem of dealing with similarity matrices containing very low values. Therefore, we believe that independently of the clustering method used, the average text length of the corpus to be clustered is an important feature that must be considered when evaluating its relative difficulty. But it is expected that after applying one enrichment mechanism, blogs corpora obtain more benefit; however, Slashdot should be the one that obtains the best performance with respect to its baseline.

B. *Stylometry-based measure*

This measure determines the language style of writing. Thus, we expect to obtain a high value when the style is very specific, whereas a low value would indicate a general language writing style. A comparison of the stylometric measures clearly indicate that the blog corpora may be documents written by many people with different writing styles. This can be contrasted with the *Cicling-2002* corpus, for instance, where we can expect a much stronger degree of

homogeneity in academic writing.

C. Class Imbalance Measure

The Class Imbalance (CI) degree of a given corpus is closely-related to the external corpus validation measure used (in our case, we use the F-measure). It gives an idea related to the number of elements in each class (twelve in the case of B-B and three for Slashdot), as to whether the corpus is balanced or unbalanced. The higher the value, the more unbalanced the corpus is. As we can see in Table V, the blog corpora appear well-balanced in comparison with the Reuters and *WSI-SemEval* corpora. We may see that both blog corpora indicate well balanced categories in the gold standard (highest values) and, therefore, this feature has neither positive nor negative impact on the clustering process.

D. Broadness Measures

Both the supervised and unsupervised measures accurately indicate that the scientific documents are narrow domain, whereas the news collections belong to a wide domain. In relation to these, the blog corpora, Slashdot with the lowest values, is shown by all measures to be of narrower domain corpus than B-B corpus is.

The Language-based approaches (SLMB and ULMB) make use of statistical language modeling in order to calculate probabilities of sequences of words (n-grams) and, thereafter, to determine the domain broadness degree of a given corpus. In the case of the vocabulary-based approaches (SVB and UVB) the measure of domain broadness assumes that corpora subsets which belong to a narrow domain share the maximum number of vocabulary terms compared with those subsets which do not. In case of wide domain corpus, it is expected (at least with short texts) that the standard deviation of vocabularies obtained from subsets of this corpus (with respect to the full corpus vocabulary) is greater than the one of a narrow domain corpus. The domain broadness corpus feature may have a strong impact in the clustering process, due to the specificity of the vocabulary.

V. Conclusions and Further Work

Characterizing the blogosphere is a new highly challenging task in the natural language processing area and can be critical in the success of the clustering task. Obtaining knowledge of the characteristics of a corpus to be analyzed can be critical to the outcome of the clustering task. The blogosphere in particular presents us with certain unique challenges, since we can expect many of its texts to be short, and of narrow domain. While this presents specific difficulties for clustering algorithms, having *a priori* characterization knowledge enables us to take remedial action by choosing the most appropriate clustering technique.

After analyzing the results of our experiments we can see what features are specific to different types of texts: scientific text, short news and blogs. In particular, we conclude that blogs were characterized as short text, with general writing style and they can be considered narrow domain due to they discuss very specific topics. Discovering automatically that some corpus is made of blogs is very important, because with this information we can design an appropriate methodology for clustering this kind of corpora [8], which takes advantage of a new self-enriching technique in order to overcome the undesirable characteristics of blogs from a clustering perspective.

Acknowledgements

The work of the first author is supported by the HEA under grant PP06TA12. The work of the second and the fourth author is supported by the TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 research project.

References

[1] A. Kilgarriff, Comparing Corpora. International Journal of Corpus Linguistics 6 (1), pages 1-37, 2001.

[2] A. Sarkar, A. De Roeck, P. H. Garthwaite, Term re-occurrence measures for analyzing style. In The SIGIR 2005 workshop on Stylistic Analysis of Text For Information Access, 2005.

[3] B. Stein and S. Meyer, F. Wißbrock, On cluster validity and the information need of users. In Proceedings of the 3rd IASTED, pages 216–221. ACTA Press, 2003.

[4] B. Stein and O. Nigemman, On the nature of structure and its identification. In Proc. of the 25th International Workshop on Graph-Theoretic Concepts in Computer Science, volume 1665 of Lecture Notes in Computer Science, pages 122–134. Springer-Verlag, 1999.

[5] C. Manning and H. Schütze, Foundations of statistical natural language processing. MIT Press, 1999.

[6] D. Perez, Evaluación de blogs, Universidad Politécnica de Valencia, Spain, Technical, Report, July 2008 (in Spanish).

[7] D. Pinto, On Clustering and Evaluation of Narrow Domain Short-Text Corpora, PhD dissertation, Universidad Politécnica de Valencia, Spain, 2008.

[8] F. Perez-Tellez, D. Pinto, J. Cardiff and P. Rosso, Improving the Clustering of Blogosphere with a Self-Term Enriching Technique, 12th International Conference on Text, Speech and Dialoge, LNCS 5729, pages 40-47, Springer-Verlag, 2009.

[9] F. Perez-Tellez, D. Pinto, J. Cardiff and P. Rosso, Characterizing Weblog Corpora, 14th International Conference on Aplications of Natural Language to Information Systems, NLDB, Springer-Verlag, 2009.

[10] G. Herdan, Quantatative Linguistics. London: Butterworth, 1964.

[11] G. K. Zipf, , Human behaviour and the principle of least effort. Addison-Wesley, 1949.

[12] H. Liu and R. Mihalcea, Of Men, Women, and Computers: Data-Driven Gender Modeling for Improved User Interfaces, in Proceedings of the International Conference on Weblogs and Social Media (ICWSM), Boulder, Colorado, March 2007.

[13] M. F. Porter, An algorithm for suffix stripping, Readings in information retrieval, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1997.

[14] R. Mihalcea and H. Liu, A corpus-based approach to finding happiness, in the AAAI Spring Symposium on Computational Approaches to Weblogs, March 2006.