Improving the Clustering of Blogosphere with a Self-term Enriching Technique

Fernando Perez-Tellez[1], David Pinto[2], John Cardiff[1], and 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

Abstract. The analysis of blogs is emerging as an exciting new area in the text processing field which attempts to harness and exploit the vast quantity of information being published by individuals. However, their particular characteristics (shortness, vocabulary size and nature, etc.) make it difficult to achieve good results using automated clustering techniques. Moreover, the fact that many blogs may be considered to be narrow domain means that exploiting external linguistic resources can have limited value. In this paper, we present a methodology to improve the performance of clustering techniques on blogs, which does not rely on external resources. Our results show that this technique can produce significant improvements in the quality of clusters produced.

1 Introduction

In recent years the use of World Wide Web has changed considerably. One of its most prominent new features is that it has become a tool in the process of socialization with services like blogs, wikis, and file sharing tools. Blogs have become a particularly important decentralized publishing medium which allows a large number of people to share their ideas and spread opinions on the internet. In order to harness the huge volume of information being published, it is essential to provide techniques which can automatically analyze and classify semantically related content, in order to help information retrieval systems to fulfill specific user needs.

The task of automatic classification of blogs is complex. Firstly, we consider it is necessary to employ clustering techniques rather than categorization, since the latter would require us to provide the number and tags of categories in advance. While we could expect to achieve better results using categorization since tags are required a priori, the dynamic categories found in blogs make clustering the best choice. Secondly, clustering techniques typically can produce better results when dealing with wide domain full-text documents where more discriminative information is available. 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

V. Matoušek and P. Mautner (Eds.): TSD 2009, LNAI 5729, pp. 40–47, 2009.

perspective such as low frequency terms, short vocabulary size and vocabulary overlapping of some domains.

In our approach, we treat blog content purely as raw text. While there exist initiatives such as the SIOC project [2] that provide the means to interconnect this information using implicit relations, they require information to be tagged in advance. By dealing purely with raw text, our approach can be applied to any blog.

The main contribution in this paper is the presentation and application of a novel approach to improve the clustering of blogs entitled "S-TEM" (Self-Term Expansion Methodology). This methodology consists of two steps. Firstly, it improves the representation of short documents by using a term enriching (or term expansion) procedure. In this particular case, external resources are not employed because we consider that it is quite difficult to identify appropriate linguistic resources for information of this kind (for instance, blogs often cover very specific topics, and the characteristics of the content may change considerably over time). Moreover, we intend to exploit intrinsic properties of the same corpus to be clustered in an unsupervised way. In other words, we take the same information that will be clustered to perform the term expansion; we called this technique self-term expansion.

The second step consists of the Term Selection Technique (TST). It selects the most important and discriminative information of each category thereby reducing the time needed for the clustering algorithm, in addition to improving the accuracy/precision of the clustering results.

Our contention is that S-TEM can help improve the automatic classification of blogs by addressing some of the factors which make it difficult to achieve strong results with clustering techniques. We show the feature analysis and clustering, using corpora extracted from the popular blogs Boing Boing ("B-B") and Slashdot[1]. In order to demonstrate that the benefits are not confined to a single clustering algorithm, we perform tests using two different clustering methods: K-star [13] and K-means [6]. The obtained results are compared and analyzed using the F-Measure, which is widely used in the clustering field [15].

The remaining of this paper is organized as follows. In the next section the corpora and preprocessing techniques are presented. In Section 3 the techniques and algorithms used in the experiments are introduced. Section 4 describes the experiments using the proposed methodology. Finally in Section 5 conclusions and future work are discussed.

2 Corpora

Table 1 presents some properties of the two datasets used in the experiments. These properties include running words (number of words in the corpus), vocabulary length, number of post categories, number of discussion lines, and the total number of posts.

The gold standard (i.e., the manually constructed expert classification) was created by Perez[2]. A discussion of the corpora is presented in [9], along with a detailed analysis of the features such as shortness degree, domain broadness, class imbalance, stylometry and structure of the corpora.

[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)

[2] Available at http://www.dsic.upv.es/grupos/nle/downloads.html (June 2009)

Table 1. Properties of the blogs datasets

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

3 Improving the Performance of the Blogosphere Clustering Task

In order to test the effectiveness of S-TEM, we have carried out a number of experiments with both corpora. Firstly we applied the methodology on the corpora, and then performed clustering on the enriched corpora. We compared the results with the clusters obtained on the original corpora (ie. without application of the enrichment methodology). For comparative purposes, we use two separate clustering algorithms, K-means [6] and K-star [13].

3.1 Description of the Clustering Algorithms

K-means [6] is one of the most popular iterative clustering algorithms, in which the number of clusters k has to be fixed a-priori. In general terms, the idea is to choose k different centroids. As different locations will produce different results, the best approach is to place these centroids as far away from each other as possible. Next, choose given data and associate each point to the nearest centroid, then k new centroids will be calculated and the process is repeated.

K-star [13] is an iterative clustering method that begins by building a similarity matrix of the documents to be clustered (corpus), In contrast with K-means, K-star does not need to know the k value a priori, and instead it automatically proposes a number of clusters in a totally unsupervised way. K-star is a considerably faster algorithm than K-means and it also obtains reasonably good results when it is applied to short text corpora. For practical purposes a minimum of document similarity threshold was established in order to permit the K-star clustering method to determine whether or not two documents belong to the same cluster. This threshold was defined as the similarity averaged among all the documents.

3.2 The Self-term Expansion Methodology

The Self-term expansion methodology ("S-TEM") [12] comprises a twofold process: the self-term expansion technique, which is a process of replacing terms with a set of co-related terms, and a Term Selection Technique with the role of identifying the relevant features.

The idea behind Term Expansion is not new; it has been studied in previous works such as [12] and [5] in which external resources have been employed. Term expansion has been used in many areas of natural language processing as in word disambiguation in [1], in which WordNet [4] is used in order to expand all the senses of a word. However, as we previously mentioned, we use only the information being clustered to perform the term expansion, i.e., no external resource is employed.

The technique consists of replacing terms of a post with a set of co-related terms. A co-occurrence list will be calculated from the target dataset by applying the Pointwise Mutual Information (*PMI*) [7] as described in Eq. (1):

$$PMI(x, y) = \log_2(N \frac{fr(x, y)}{fr(x) \cdot fr(y)}), \quad (1)$$

where *fr(x,y)* is the frequency in which both words *x* and *y* appear together; *fr(y)* and *fr(x)* are the word frequency of *x* and *y*, respectively, and *N* is a normalization factor equal to the total number of words in the vocabulary.

PMI provides a value of relationship between two words; however, the level of this relationship must be empirically adjusted for each task. In this work, we found *PMI* equal or greater than 2 to be the best threshold. This threshold was established by experience, analyzing the performance of clustering algorithms with different samples of the datasets. In other experiments [11], [12], a threshold equal to 6 was used; however, in the case of blog corpora documents, correlated terms are rarely found.

The Self-Term Expansion Technique is defined formally in [10] as follows:

Let $D = \{d_1, d_2, \ldots, d_n\}$ be a document collection with vocabulary *V(D)*. Let us consider a subset of $V(D) \times V(D)$ of co-related terms as $RT = \{(t_i, t_j) | t_i, t_j \in V(D)\}$ *The RT* expansion of *D* is $D' = \{d'_1, d'_2, \ldots, d'_n\}$, such that for all $d_i \in D$, it satisfies two properties: 1) if $t_j \in d_i$ then $t_j \in d'_i$, and 2) if $t_j \in d_i$ then $t'_j \in d'_i$, with $(t_j, t'_j) \in RT$. If *RT* is calculated by using the same target dataset, then we say that *D'* is the self-term expansion version of *D*.

The Term Selection Technique helps us to identify the best features for the clustering process. However, it is also useful to reduce the computing time of the clustering algorithms. In particular, we have used Document Frequency (DF) [14], which assigns the value DF(*t*) to each term *t*, where DF(*t*) means the number of posts in a collection, where *t* occurs.

The Document Frequency technique assumes that low frequency terms will rarely appear in other documents; therefore, they will not have significance on the prediction of the class of a document. This assumption is completely valid for all textual datasets, including blog corpora.

4 Experimental Results

In this section the performance of the two clustering algorithms in conjunction with S-TEM is presented. The goal is to show one strategy that could be used in order to deal with some problems related with blogs such as, low frequency terms, short vocabulary size and vocabulary overlapping of some domains.

We apply S-TEM in order to replace terms in blogs with a list of co-related terms; this list may be obtained by using general purpose and external knowledge resources; however, due to the topic specificity of blogs, there is a lack of linguistic resources of this kind. Intrinsic information in the target dataset should be exploited together with a selection of terms in order to use the most important and relevant information needed for the clustering task.

The results we obtained with and without the Self-Term Expansion Methodology can give us an overview of the level of improvement that may be obtained by applying this methodology. In order to be objective with the results, we have used

a well-known measure to evaluate the performance of the clustering algorithms, which is named F-Measure [15] and it is described as follows:

$$F - Measure = \frac{2 * precision * recall}{precision + recall} \quad (2)$$

The clustering results of applying S-TEM to the B-B and Slashdot corpora are shown in Figures 1 and 2 respectively. Different vocabulary percentages of the enriched corpus were selected by the Term Selection Technique (from 30% to 90%).

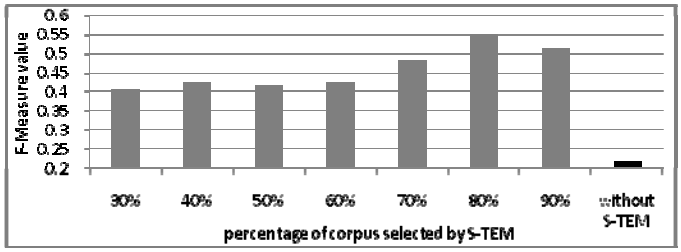

Fig. 1. Applying S-TEM to B-B and clustering with K-means

In general the best clustering results were obtained by selecting (using the DF technique) 80% of the total number of terms belonging to the B-B enriched corpus vocabulary, and using the value of *k=13* in K-means algorithm; this parameter *k* (number of clusters) was estimated by experience checking the best performance of the clustering algorithm, whereas with Slashdot in Figure 2, the best result was obtained selecting 40% of its vocabulary.

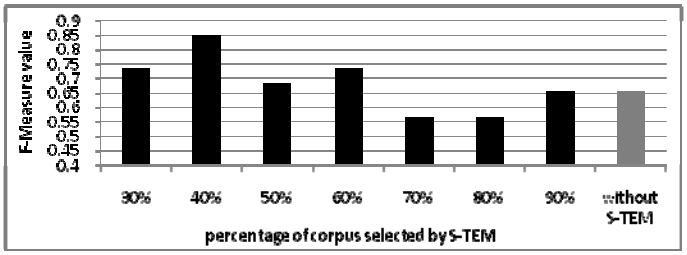

Fig. 2. Applying S-TEM to Slashdot and clustering with K-means

In the case of B-B (Figure 1), the quality of clusters shows a significant improvement when S-TEM is applied. The enriched corpus obtained a F-Measure of 0.55, whereas the baseline (non-enriched) version obtained F-Measure equal to 0.25.These results confirm the usefulness of the methodology with short-text and narrow domain corpora.

The clustering results applied to the Slashdot corpus are shown in Figure 2. Again, the impact of the S-TEM in the clustering of Slashdot produces a considerable improvement in the quality of results, even though the number of documents of the corpus is considerably smaller than B-B. The best F-Measure value (0.85) is obtained

with 40% of the vocabulary of the enriched corpus, and *k* of K-means equal to five; this parameter was estimated in the same form as in B-B. In contrast with the best value obtained by K-means without using S-TEM that is 0.65.

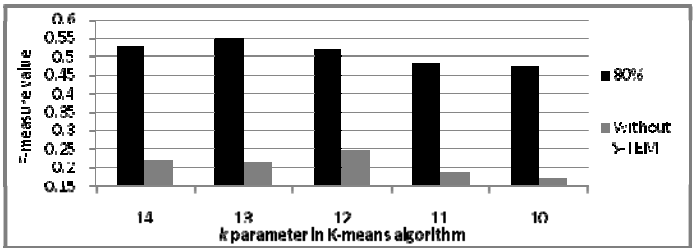

Fig. 3. Varying *k* parameter in K-means for B-B corpus

In Figures 3 and 4 the results using different values for the *k* parameter in the K-means algorithm are presented (from 10 to 14 for B-B and from 2 to 7 for Slashdot). The importance of varying the value of *k* is to confirm the number of clusters that can produce the best results after applying S-TEM to the dataset which will be clustered.

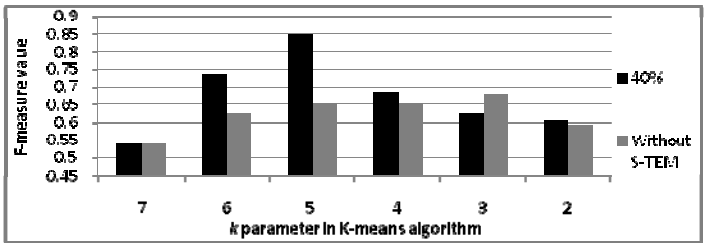

Fig. 4. Varying *k* parameter in K-means for Slashdot corpus

Additionally, Figure 3 shows a considerable improvement obtained with *k=13* by the clustering algorithm using S-TEM, whereas in Figure 4 the best value was with *k* =5. The improvement in the results is smaller but still considerable, due to the number of documents and discriminative information used by the K-means algorithm.

In order to more easily understand the obtained results, we have calculated the percentage growth of the enriched corpora with respect to the original data. In Table 2 we see that B-B is the best benefited with a very high percentage of enriching terms. We consider this fact was fundamental for the improvement obtained by using S-TEM, when the results are compared to the baselines.

Table 2. Increase of vocabulary after applying the Self-term Expansion Technique

Boing Boing	Slashdot
+ 1665.8 %	+ 70.29 %

In order to analyze the performance of S-TEM in blogs with different kind of clustering algorithms, we compare the performance improvement of both corpora with the K-star clustering algorithm before and after applying S-TEM. We focused our analysis with the 80% of the enriched corpus vocabulary. As shown in Table 3, when using S-TEM, the improvement is more than 100%. We have confirmed the results obtained with K-means, although we have used another paradigm of clustering, where the number of clusters is automatically discovered. K-star identified nine of the twelve real categories of B-B when the S-TEM was applied; however, 30 categories were discovered when S-TEM was not applied.

Table 3. F-Measure values of the K-star algorithm

Corpus	No. of classes found by K-star without S-TEM	Number of classes found by K-star with S-TEM	F-Measure Without S-TEM	F-Measure With S-TEM
B-B	30	9	0.2	0.41
Slashdot	12	2	0.44	0.45

The values shown in Table 3 by the K-star when clustering the Slashdot corpus with and without S-TEM are quite similar. We consider this is the result of the poor vocabulary of this particular corpus, in addition to noisy words and high overlapping vocabulary between classes. These issues have a notable effect on the clustering task. The K-star algorithm suggested that only two clusters should be considered in the Slashdot corpus when S-TEM was used, however, twelve clusters were identified when S-TEM was not used.

K-means obtained better results for both corpora than K-star. However, we must consider that K-means is computationally more expensive and it requires us to fix the value of *k* a priori, whereas K-star is a completely unsupervised method.

5 Conclusions and Further Work

Clustering of blogs is a highly challenging task. Blogs are often characterized as narrow domain short-texts and, are therefore unsuitable for augmentation by using external resources. In this paper, we presented a text enrichment technique, the aim of which is to improve the quality of the corpora with respect to the clustering task. The novel feature of this approach, the Self-Term Expansion Methodology (S-TEM) is that it does not rely on external linguistic resources, but it uses the corpus to be clustered itself. We presented a comparison of using the S-TEM with two blog datasets. The results obtained in the experiments show a significant improvement of the enriched corpora with respect to its baseline (non-enriched version).

In future work we plan to use other co-relation measures like those described in [3] and [8], which will be applied in order to find better relationships between terms, with the goal of providing information which is beyond the lexical level to the clustering algorithms.

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. The authors wish to express their thanks to Prof. Mikhail Alexandrov for his invaluable feedback on this research.

References

1. Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordNet. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 136–145. Springer, Heidelberg (2002)
2. Bojars, U., Breslin, J.G., Passant, A.: SIOC Browser Towards a richer blog browsing experience. In: The 4th BlogTalk Conference (2006)
3. Daille, B.: Qualitative terminology extraction. In: Bourigault, D., Jacquemin, C., et l'homme, M.-C. (eds.) Recent Advances in Computational Terminology. Natural Language Processing, vol. 2, pp. 149–166. John Benjamins (2001)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
5. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Academic Publishers, Dordrecht (1994)
6. MacQueen, J.B.: Some methods for classification and analysis of multivariate observations. In: Proc. of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
7. Manning, D.C., Schütze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge (1999)
8. Nakagawa, H., Mori, T.: A Simple but Powerful Automatic Term Extraction Method, International Conference on Computational Linguistics. In: COLING 2002 on COMPUTERM 2002: second international workshop on computational terminology, vol. 14 (2002)
9. Perez-Tellez, F., Pinto, D., Rosso, P., Cardiff, J.: Characterizing Weblog Corpora. In: 14th International Conference on Applications of Natural Language to Information Systems (2009)
10. Pinto, D., Rosso, P., Jiménez-Salazar, H.: UPV-SI: Word Sense Induction using Self-Term Expansion. In: 4th Workshop on Semantic Evaluations - SemEval 2007, Association for Computational Linguistics (2007)
11. Pinto, D.: On Clustering and Evaluation of Narrow Domain Short-Text Corpora, PhD dissertation, Universidad Politécnica de Valencia, Spain (2008)
12. Qiu, Y., Frei, H.P.: Concept based query expansion. In: Proc. of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 160–169. ACM Press, New York (1993)
13. Shin, K., Han, S.Y.: Fast clustering algorithm for information organization. In: Gelbukh, A. (ed.) CICLing 2003. LNCS, vol. 2588, pp. 619–622. Springer, Heidelberg (2003)
14. Spärck, J.K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)
15. Van Rijsbergen, C.J.: Information Retireval. Butterworths, London (1979)