Clustering Weblogs on the Basis of a Topic Detection Method

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
fernandopt@gmail.com, John.Cardiff@ittdublin.ie
[2] Benemérita Universidad Autónoma de Puebla, Mexico
dpinto@cs.buap.mx
[3] Natural Language Engineering Lab, ELiRF, Universidad Pólitecnica de Valencia, Spain
prosso@dsic.upv.es

Abstract. In recent years we have seen a vast increase in the volume of information published on weblog sites and also the creation of new web technologies where people discuss actual events. The need for automatic tools to organize this massive amount of information is clear, but the particular characteristics of weblogs such as shortness and overlapping vocabulary make this task difficult. In this work, we present a novel methodology to cluster weblog posts according to the topics discussed therein. This methodology is based on a generative probabilistic model in conjunction with a Self-Term Expansion methodology. We present our results which demonstrate a considerable improvement over the baseline.

Keywords: Clustering, Weblogs, Topic Detection.

1 Introduction

In recent years the World Wide Web has shown huge changes as a tool of socialization, bringing up new services and applications such as weblogs, wikis as part of the Web 2.0 technologies. The blogosphere is a new medium of expression, becoming more popular all around the world. We can find weblogs in all subjects from sports, games to politics and finance.

In order to manage the large amount of information published in the blogosphere, there is a clear need for systems that provide automatic organization of its content, in order to exploit the information more efficiently and retrieve only the information required for a particular user. Document clustering –the assignment of documents to previously unknown categories— has been used for this purpose [20]. We consider it more appropriate to employ clustering rather than classification, since the latter would require providing tags of categories in advance and in real scenarios we usually deal with information from the blogosphere without knowing the correct category tag.

The focus of this research work is to study a novel approach for clustering weblog posts according to their topics of discussion. For this purpose, we have based our approach in a topic detection method. Topic detection and tracking is a well-studied

J.A. Carrasco-Ochoa et al. (Eds.): MCPR 2010, LNCS 6256, pp. 342–351, 2010.

area [2] [3], which focuses on extraction of significant topics and events from news articles. We consider the topic detection task as the problem of finding the most prominent topics in a collection of documents; in general terms, identifying a set of words that constitute topics in a collection of documents.

The main contribution in this work is a novel methodology of clustering weblog posts based on a topic detection model for text in conjunction with a Self-Term Expansion methodology [16]. In our approach we treat the weblog content purely as raw text, identifying the different topics inside of the documents and using this information in the clustering process.

In [15], the features of weblogs are discussed, for instance, weblogs can be characterized as very short texts and with a general writing style. These are undesirable characteristics from a clustering perspective, as not enough discriminative information is provided. In order to tackle the particular characteristics of weblogs, we employ an expansion methodology, the Self-Term Expansion Methodology [16], that does not use external resources, relying only on information included in the corpus itself then. Our hypothesis states that the application of this methodology can improve the quality of topic clusters, and further that the improvement will be more significant where the corpus is composed of well-delimited categories which share a low percentage of vocabulary (wide domain corpus).

The methodology we present consists of four parts. Firstly, it improves the representation of the text by means of a Self-Term Enriching Technique. External resources are not employed because we consider it difficult to identify appropriate linguistic resources for information such weblogs. Secondly, a Term Selection Technique is applied in order to select the most important and discriminative information of each category thereby reducing processing time for the next two steps. The third step is the use of the Latent Dirichlet Allocation method [5], which is a generative probabilistic model for discrete data. We use this model to construct a set of reference vectors which can be used as categories prototypes for a better and faster clustering process. Finally, we use the well-known Jaccard coefficient [14] as a similarity measure to form the clusters.

The rest of this paper is organized as follows. Section 2 presents the related work. Section 3 describes the dataset used in the experiments. Section 4 explains our approach and the techniques used in our research work. Section 5 shows the obtained results. Section 6 provides an analysis of results and, finally, in Section 7 we present the conclusions.

2 Related Work

There are previous attempts on topic detection in online documents such as in [8], where the authors present a topic detection system composed of three modules that attempt to model events and reportage in news. The first module (pre-processing) is used to select and weight the features, i.e., words that are representative of short events. The clustering module is a hybrid technique that uses a slow accurate hierarchical method with a fast partitional algorithm. Finally, the last module is the presentation module which displays each cluster to the user.

The task of finding a set of topic in a collection of documents has also been attempted in [21]; the authors based their approach on the identification of clusters of keywords that are taken as representation of topics. They have employed the well-known k-means algorithm to test some distance measures based on a distribution of words. The experiments were conducted using Wikipedia articles, reporting acceptable results, but the calculation of the distributions seems to be computational expensive.

Topic detection is also addressed in [18], where the authors present a method which uses blogger's interests in order to extract topic words from weblogs. In this approach the authors assume that topic words are words commonly used by bloggers who share the same interests, and they use these topic words to compute similar interests between each two bloggers by using the cosine similarity measure. A topic score is assigned to each word. The processing time is also a problem in this approach, as they have pointed out, and the optimization for some of their calculations is needed.

Recently, the clustering of weblogs has become an active topic of research; for instance in [13] the authors build a word-page matrix by downloading weblog pages and have applied the k-means clustering algorithm with different weights assigned to the title, body, and comment parts. In [1], the authors use weblog categories to build a category relation graph in order to join different categories; they use edges in the category relation graph to represent similarity between different categories and they represent nodes as categories. They also consider different values of link strengths and level of directories.

Our approach is focused on detecting the topic clusters contained in the corpus itself, and the novel aspect is based on using a topic detection method to identify possible references that could be used in the clustering process, and the expansion methodology in order to improve the representation of the weblogs.

3 Description of Dataset

In this section, we describe the corpus used in our experiments. The corpus is a subset of the ICWSM 2009 Spinn3r Blog Dataset[1], the content of the data includes metadata such as the blog's homepage, timestamps, etc. The data is in XML format and according to the Spinn3r crawling[2] documentation; it is further arranged into tiers, approximating search engine ranking to some degree.

Even if the Spinn3r blog dataset contains several blogs sites in a number of different languages, we only focused the experiments carried out on the "Yahoo Answers", weblog site[3] – in which people share what they know and ask questions on any topic that matters to the user, in order to be answered by other users. We have extracted from this corpus two distinct subsets (see Fig. 1). The first subset contains 10 categories with 25,596 posts and vocabulary size of 66,729. It may be considered as "narrow domain", since the vocabulary in the categories is similar. The second

[1] The corpus was initially made available for the 2009 Data Challenge at the 3rd International AAAI Conference on Weblogs and Social Media, http://www.icwsm.org/2009/data/

[2] http://spinn3r.com/documentation/

[3] http://answers.yahoo.com/

subset contains 10 categories with 48,477 posts and a vocabulary size of 122,960 terms. As opposed to the narrow domain subset, it may be considered "wide domain" because its categories have a low overlapping vocabulary.

	Category name	Posts	Category name	Posts
Subset 1 (Narrow Domain)	Cell_Phones_Plans	1,543	Video_Online_Games	6,578
	Computer_Networking	1,337	Maintenance_Repairs	1,973
	Programming_Design	2,466	Security	1,583
	Laptops_Notebooks	2,153	Music_Music_Players	1,640
	Software	4,800	Other_-_Internet	1,523
Subset 2 (Wide Domain)	Singles_Dating	20,498	Celebrities	2,219
	Software	4,800	Marriage_Divorce	2,956
	Womens_Health	4,262	Languages	1,914
	Politics	2,527	Elections	3,628
	Dogs	3,205	Books_Authors	2,468

Fig. 1. Topics of discussion of the two datasets (narrow and wide domain)

Clustering of narrow domains brings additional challenges to the clustering process. Moreover, the shortness of this kind of data will make this task more difficult. The purpose of constructing two subsets with these characteristics is to demonstrate the effectiveness of our method across both wide and narrow domains, and also to test the relative effectiveness of the approach in each case.

Regarding the categories tags, they were only used for gold standard construction purposes, and provide a better idea of the subsets used in our experiments. The posts are treated as raw text, i.e. we have not used any additional information provided by the XML tags. As a preprocessing step, we have removed stop words –high-frequency word that has not significant meaning in a phrase– and punctuation symbols as well.

4 Methodology Proposed

In this section, we present the techniques used in our approach in order to improve the quality of clusters. This methodology clusters weblog posts using prototypes as reference, therefore, we have also called this approach prototype/topic based clustering. Our approach is composed of three steps: the Self-Term Expansion Methodology (S-TEM), which consists of a Self-Term Enriching Technique and a Term Selection Technique. This is followed by the application of the Latent Dirichlet Allocation model and the prototype/topic based clustering process.

4.1 Self-Term Expansion Methodology

The Self-Term Expansion Methodology [16] comprises a twofold process: the Self-Term Enriching 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 has been studied in previous works such as [17] and [9] in which external resources have been employed. Term expansion has been used in many areas of natural language processing as in word disambiguation in [4], in which WordNet [7] is used in order to expand all the senses of a word. However, in the particular case of the S-TEM methodology, 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 web post with a set of co-related terms. We consider it particularly important to use the intrinsic information of the data set itself. A co-occurrence list is calculated from the target dataset by applying the Pointwise Mutual Information (*PMI*) [14]. *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 3 to be the best threshold. This threshold was established empirically. In other experiments [16], a threshold of 6 was used; however, in weblog documents correlated terms are rarely found. This list will be used to expand every term of the original corpus.

The Self-Term Enriching Technique is defined formally in [16] 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 degree of co-occurrence between a pair of terms is determined by a co-ocurrence method, this method is based on the assumption that two words are semantically similar if they occur in similar contexts [10].

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) [19], 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.

4.2 Latent Dirichlet Allocation Model

In general, a topic model is a hierarchical Bayesian model that associates each document to a probability distribution over topics. The *Latent Dirichlet Allocation* (LDA) model [5] is derived from the idea of discovering short descriptions of the members of a collection, in particular discrete data, in order to allow efficient processing of huge collections, while keeping the essential statistical relationships that may be used in other tasks such as classification.

There are other sophisticated approaches that use dimensionality reduction techniques such as *Latent Semantic Indexing* (LSI) [6], which can achieve significant compression in large corpora using single value decomposition of the X matrix to identify a linear subspace in the space of *tf-idf* features by capturing most of the variance in the corpora. An alternative model is *probabilistic Latent Semantic Index* (pLSI) [11], in which the main idea is to model each word in a document as a sample

from a mixture model, in which the components of the mixture are multinomial random variables that can be viewed as words generated from topics. However, LDA may be seen as a step forward with respect to LSI and pLSI.

The LDA model is based on a supposition that the words of each document arise from a mixture of topics, each of that is a distribution over the vocabulary. This method has been used for automatically extracting the topical structure of large document collections, in other words, it is a generative probabilistic model of a corpus that uses different distributions over a vocabulary in order to describe the document collection.

4.3 Clustering Weblog Posts Using the Prototypes as References

The prototype/topic based clustering methodology is outlined in Fig. 2. We start from having the corpus as raw text. Then we apply the S-TEM approach to the original posts. In the Term Selection Technique we have selected from 10% to 90% of vocabulary after the enriching process, in order to confirm which percentage provides the best information to LDA Method.

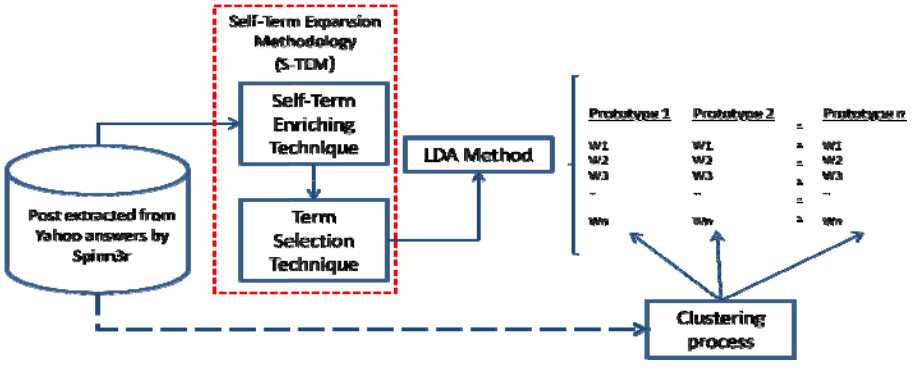

Fig. 2. Methodology proposed "prototype/topic based clustering"

The LDA method generates the prototypes, i.e., vectors that will contain topics discussed on the posts. We expect to have a reference for each category in order to generate the clusters, one for each prototype. In this step, LDA requires as input the number of possible topics, in our case we have fixed this parameter to ten, which is the number of categories in each subset. We have also varied the number of terms selected from 100 to 3,000 in order to confirm the best and minimum number of terms for the clustering task.

Finally, the clustering process will compare each original post (unexpanded) with each prototype; every post will be assigned to one cluster according to the most similar prototype (highest value in the clustering process). We have chosen the Jaccard coefficient because its simplicity and relative fast clustering process. In our case, we have compared each original post against each prototype and the highest similarity measure with the prototypes get the post in its cluster.

5 Experiments

In this section, we present the experiments and results using the approach proposed in this research work. These experiments were carried out over the two subsets described in Section 3.

5.1 Wide Domain Subset

Fig. 3 presents a comparison of our approach against the baseline for the wide domain corpus. We have obtained the baseline by generating the prototypes with the LDA method from the original posts, i.e., without using the S-TEM methodology in the construction of the prototypes, and finally, clustering the posts with the Jaccard coefficient. We have summarized the results in the graph showing the minimum, maximum and average F-measure value obtained from the different percentage of vocabulary selected (from 10% to 90% with steps of 10%) with the Term Selection Technique in the S-TEM methodology.

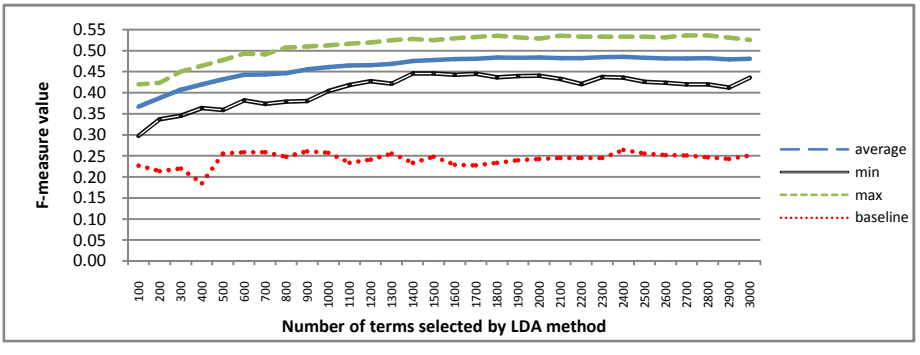

Fig. 3. Clustering results using the "wide" domain corpus

The objective of using this selection is to reduce the noise (terms included in more than one category that can be highly correlated with discriminative information) generated by the enriching technique and to highlight the most important features of each category. We have obtained the best results when we have selected 10% of vocabulary (achieving an F-measure value of 0.53). It means that after the enriching process, it only needs 10% of the vocabulary to generate the best prototypes. We have also confirmed that in all the cases we have outperformed the baseline (0.26 in the best case). We have limited the number of terms selected by the LDA method from 100 to 3,000 terms per topic in order to confirm the minimum number of terms for the prototype which can give us acceptable results in the clustering process. Furthermore, by reducing the number of terms, we can reduce the processing time for the clustering task.

5.2 Narrow Domain Subset

In Fig. 4 we present the improvement that the S-TEM methodology provides to this clustering approach for the narrow domain corpus. In this particular case the gap between the baseline and the average is smaller.

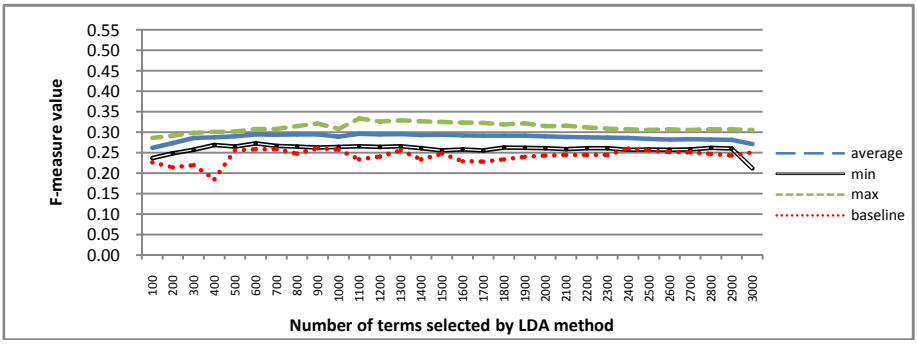

Fig. 4. Clustering results using the "narrow" domain corpus

In other words, the performance of our methodology is not as high as that obtained with wide domain, but in any case we still achieve an improvement. We consider that the reduced improvement in this domain is due to the fact that when the enrichment process expands the corpus, it introduces some noisy terms, i.e., terms that share many categories in this kind of domain. Even if we have used the Term Selection Technique to avoid this noisy information, it is difficult to highlight the discriminative information of each category. All of this makes the clustering task more difficult. Therefore, the size of the each document (in this case, weblog posts) is another important factor involved in this complex clustering process.

6 Analysis of Results

In this section, we discuss the results obtained in the experiments. As we expected we have obtained the best results with the wide domain corpus, because the categories share a low percentage of vocabulary. On the other hand, the narrow domain has a very high overlapping vocabulary between categories, which is a very important factor reflected in the clustering process. We have found out that the S-TEM methodology can help the generation of prototypes because the LDA has taken advantage of the expansion methodology. The improvement of the representation that S-TEM gives to the narrow domain posts is less because of the high overlapping vocabulary, and also the noise introduced by the enriching process derived from the Pointwise Mutual Information that is based on the frequency of correlated terms. It is also important to mention that we have outperformed the baseline in both cases (narrow and wide domain).

An additional aspect found in our experiment and shown in Figures 3 and 4 is that using nearly a thousand terms per category in the prototypes is good enough to get acceptable result the clustering process this may impact in the processing time due to we can manage relatively low-dimension vectors.

7 Conclusions and Further Work

We have presented a novel methodology to cluster weblogs based on a generative probabilistic model (LDA) in conjunction with an enriching methodology (S-TEM)

applied to two different kind of corpus, one considered as “narrow” domain with very similar categories, and other considered as “wide” domain with low overlapping vocabulary or dissimilar categories.

We have confirmed that our approach works well with wide domain corpora obtaining 0.53 in F-measure with just 10% of the vocabulary to generate the best prototypes and it has also shown improved results (albeit with a smaller gain) with narrow domains. Finally, due to the simplicity of the clustering method used, our approach has shown acceptable ranges in the processing time.

In future work, we plan to modify our approach and cluster the expanded posts used in the generation of the prototypes with the objective of giving better information to the clustering process and improve representation of the post in particular in narrow domain. We are also interested in working on the scalability of our approach in order to be able to manage data sets with huge number of documents and classes. To further this aim, we are intending to adapt the approach described in [12].

Acknowledgments. The work of the fourth author has been partially supported by the TEXTENTERPRISE 2.0 TIN2009-13391-C04-03 research project and the work of the first author by the Mexican Council of Science and Technology (CONACYT)**.**

References

1. Agrawal, N., Galan, M., Liu, H., Subramanya, S.: Clustering blogs with collective wisdom. In: Proc. of the International Conference on Web Engineering, pp. 336–339. IEEE Computer Society, USA (2008)
2. Allan, J., Carbonell, J.G., Doddington, G., Yamron, J., Yang, Y.: Topic Detection and Tracking Pilot Study: Final Report. In: Proc. DARPA Broadcast News Transcription and Understanding Workshop (1998)
3. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proc. SIGIR International Conference on Research and Development in Information Retrieval, pp. 37–45. ACM, NY (1998)
4. Banerjee, S., Pedersen, T.: An adapted Lesk algorithm for word sense disambiguation using WordNet. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 136–145. Springer, Heidelberg (2006)
5. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. The Journal of Marchine Learning Research, JMLR.org 3, 993–1022 (2003)
6. Deerwester, S., Dumais, S., Landauer, T., Furnas, G., Harshman, R.: Indexing by latent semantic analysis. Journal of American Society of Information Science 41, 391–407 (1990)
7. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
8. Flynn, C., Dunnion, J.: Topic Detection in the News Domain. In: Proc. of the 2004 International Symposium on Information and Communication Technologies, pp. 103–108. ACM, New York (2004)
9. Grefenstette, G.: Explorations in Automatic Thesaurus Discovery. Kluwer Ac., Dordrecht (1994)
10. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)

11. Hofman, T.: Probabilistic latent semantic indexing. In: Proc. of the Twenty-Second Annual International SIGIR Conference, pp. 50–57. ACM, NY (1999)
12. Karp, R.M., Rabin, M.O.: Efficient Randomized Pattern-Matching Algorithms. IBM Journal of Research and Development 31(2), 249–260 (1987)
13. Li, B., Xu, S., Zhang, J.: Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments. In: ACM Southeast Regional Conference, pp. 94–99 (2007)
14. Manning, D.C., Schutze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge (1999)
15. Perez-Tellez, F., Pinto, D., Cardiff, J., Rosso, P.: Characterizing Weblog Corpora. In: Horacek, H., Métais, E., Muñoz, R., Wolska, M. (eds.) NLPIS 2010. LNCS, vol. 5723, pp. 299–300. Springer, Heidelberg (2010)
16. Pinto, D.: On Clustering and Evaluation of Narrow Domain Short-Text Corpora. PhD dissertation, Universidad Politecnica de Valencia, Spain (2008)
17. 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, New York (1993)
18. Sekiguchi, Y., Kawashima, H., Okuda, H., Oku, M.: Topic Detection from Blog Documents Using Users' Interests. In: Proc. of the 7th International Conference on Mobile Data Management (2006)
19. Spärck, J.K.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11–21 (1972)
20. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: KDD Workshop on Text Mining (2000)
21. Wartena, C., Brussee, R.: Topic Detection by Clustering Keywords. In: Proc. of the 19th International Conference on Database and Expert Systems Application, pp. 54–58. IEEE Computer Society, USA (2008)