A Methodology to Cluster Informal Language Register Data

Fernando Perez-Tellez[1], David Pinto[2], John Cardiff[1], 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
[3]Natural Language Engineering Lab. – EliRF, Dept. Sistemas Informáticos y Computación, Universidad Politécnica Valencia, Spain

Abstract. Analyzing and classifying web content is a task that has been attracting an increasing amount of interest in recent years. However there are additional challenges to face with user generated content emanating from Web 2.0 applications such as blogs, commentaries, reviews etc. The typical characteristics of this information include features such as shortness, overlapping vocabulary, and vocabulary size and nature that make it difficult to achieve good results using automated clustering processes. The Web 2.0 informal written register introduces further challenges, containing incomplete sentences, misspellings, spontaneous structures etc. These characteristics make it difficult to select or employ external resources to improve the clustering performance. In this work we apply a methodology that does not rely on any external resources in order to automatically cluster this data. This approach improves the representation of informal language register data by using a term enriching procedure and also uses a term selection technique to identify the most important and discriminative information .Our results show that this technique can produce significant improvements in the quality of clusters produced.

Keywords: Blogs, Clustering, Web 2.0

1 Introduction

In recent years, Web 2.0 applications have been gaining increasing importance as a new medium for communication and publication. This has been accompanied by a corresponding interest in developing computational approaches to web content analysis for a wide variety of purposes including marketing strategies [7] [11], opinion analysis [5] [2] and education [12]. This data can be typically found in weblogs or web forums in which people spread their opinions and ideas. The language used in these cases can be classified as informal language register: it has until recent years been confined largely to speech and new techniques for analysis and

clustering of this data are required to deal with its specific characteristics now that vast quantities are widely available in written form.

In our approach, we treat informal language register data purely as raw text i.e. we discarded all the XML tags contained in the original corpus, in order not to depend on extra information and propose a more general methodology that may be used in other scenarios. There exists an approach described in [1], in which the authors deal with clustering short text, in particular scientific text, in which the authors use an open access to document images (vectors of words frequencies, which can be restricted to a small list of keywords) of the papers in order to facilitate the clustering in the search task. Another approach [4] uses verb class information and external online resources to automatically classify blog sentiment. An approach to improving the quality of clustering is presented in [20] in which the authors apply different weightings to the blog title, body, and comments. In [13], a probabilistic graphical model is proposed for grouping blog entries into temporal discussions using query keywords.

The main contribution in this paper is the presentation and application of an approach to improve the clustering of informal language register data: a text enrichment methodology based on self-term expansion [17]. This methodology consists of two steps. In the first step, we improve 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 a short space of time). Instead, we 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. It selects the most important and discriminative information of each category and, therefore, it reduces the time needed for the clustering algorithm in addition to improving the accuracy/precision of the clustering results.

The remainder of this paper is organized as follows. In the next section the corpora and the preprocessing we perform are presented. In Section 3 we describe the methodology and techniques used in our experiments. In Section 4, we analyze the outcomes of the experiments. Finally in Section 5 conclusions and future work are discussed.

2 Corpus

In this section, we describe the corpus used for our experiments, and discuss the preprocessing and extraction techniques we employed to construct a subset appropriate for testing our hypotheses. The corpus is a subset of the ICWSM 2009

Spinn3r Blog Dataset[1], a collection of 44 million blog posts provided by Spinn3r.com[2], made between August 1st and October 1st, 2008. The content of the data includes metadata such as the blog's homepage, timestamps, etc. The data is in XML format and according to Spinn3r crawling[3] documentation; it is further arranged into tiers, approximating search engine ranking to some degree. It contains information in several different languages but in our case we consider only information in the English and Spanish languages[4].

2.1 Subset Extraction and Preprocessing

Table 1 shows properties including running words (number of words in the corpus) vocabulary size and number of categories. We selected four popular post categories from the corpus to perform the experiments (Technology, Politics, Music, Sports for English and the equivalent categories of Tecnología, Política, Musica, Deportes for Spanish). In order to reduce the processing time, we selected 500 posts randomly from each category, giving a total of two thousand documents each for English and the same number for Spanish.

Table 1. Properties of the subset used.

Language	Running words	Vocabulary size	Categories
EN	331,638	39,969	4
SP	427,199	71,024	4

Figure 1 shows an example of the original corpus. The first step in constructing the corpus was to identify a document delimiter as many documents are contained in the same file. These were identified using the *"<item>"* and *"</item>"* tags. Next we selected all the information in English or Spanish, using the Dublin Core[5] metadata element *"<dc:lang>"* to identify the language. A large percentage of the corpus posts use the Dublin Core ontology for such high level descriptions thereby providing a useful means of identifying the language. Next, we filtered the information classified into the four above mentioned categories. Finally, the information included between the tags *"<description>"* and *"</description>"* was used for the construction of our corpus.

In Figure 1 is possible to identify the tags used in the construction of this corpus. Before applying the clustering algorithms, it was necessary to preprocess the corpus in order to select the most important features of each class.

[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://blog.spinn3r.com/crawler/

[4] The corpus will be made available upon request by email to the first author.

[5] http://dublincore.org/

<item>
<title>Distinguishing</title>
<link>http://ceruleanbill.blogspot.com/2008/08/distinguishing.html</link>
<guid>http://ceruleanbill.blogspot.com/2008/08/distinguishing.html</guid>
<pubDate>Fri, 01 Aug 2008 12:47:31 GMT</pubDate>
<dc:source>http://ceruleanbill.blogspot.com</dc:source>
<weblog:title>Bill's Stuff</weblog:title>
<weblog:description>Oui, Nous Pouvons Ja K¶nnen Wir - YES, WE CAN - S¬, Possiamo S, Podemos Sim, N³s Podemos</weblog:description>
<dc:lang>en</dc:lang>
<weblog:tier>205</weblog:tier>
<atom:author>
<atom:name>Cerulean Bill</atom:name>
<atom:email>noreply@blogger.com</atom:email>
<atom:link>http://www.blogger.com/profile/01212953653794261 33</atom:link>
</atom:author>
<weblog:indegree>3</weblog:indegree> <weblog:iranking>0</weblog:iranking>
<category>Politics</category>
<description>I don't know what's going on with McCain and Obama's ads -- I think McCain's is stupid, which I assume his people are not, and I think Obama's is about race, though he denies it; in both cases, I assume it's an attempt to deflect their opponent's attention to trivialities -- but I learned something this morning, from the LA Times -- Just for a start, industry types say the (McCain) ad is wrong
</description>
<weblog:publisher_type>WEBLOG</weblog:publisher_type>
<atom:published>2008-08-01T12:15:00Z</atom:published>
<post:date_found>2008-08-01T12:47:31Z</post:date_found>
<post:resource_guid>wz~A9fm2fpY</post:resource_guid>
<source:resource>http://ceruleanbill.blogspot.com</source:resource>
<post:spam_probability>0</post:spam_probability> <source:title>Bill's Stuff</source:title>
<source:description>Oui, Nous Pouvons Ja K¶nnen Wir - YES, WE CAN - S¬, Possiamo S, Podemos Sim, N³s Podemos</source:description><source:tier>205</source:tier>
<source:indegree>3</source:indegree> <source:iranking>0</source:iranking>
</item>

Fig. 1. Example of original dataset.

The preprocessing task consisted of two steps:

- *Cleaning*: All non ASCII symbols are removed from the corpus. Moreover, all characters are transformed to lower case.
- *Stemming*: We apply the stemming algorithm [19] [22], which reduces a word to its root. This process is performed with the aim of increasing recall.

Figure 2 shows a preprocessed version of the corpus. The document identifier (in bold) is shown with the category and number of the document. This identifier will be used for the evaluation measures in the following section.

politics2 south africa should be worri about the implic of jacob zuma constitut court rule sai the mkhonto we sizw veteran associ
politics3 i dont know what go on with mccain and obama ad i think mccain is stupid which i assum hi peopl ar not and i think obama is about race though he deni it in both case i assum it an attempt to deflect their oppon attent to trivial but i learn someth thi morn from the la time just for a start industri type sai the mccain ad is wrong
politics4 the other dai the new york time distanc itself from the republican it endors a half year ago in the new york primari when it critic republican sen john mccain for run the low road express in recent week mr mccain ha been wave the flag of fear senat barack obama want to lose in iraq and issu attack that ar sophomor suggest that mr obama is a socialist and fals the presumpt democrat nomine turn hi back on wound soldier to what degre these attack or critic were unfair im not go to address but i notic hypocrisi in

Fig. 2. Dataset preprocessed (extract).

2.2 Gold Standard Construction

In order to evaluate the results of the clustering task, we needed to construct a gold standard describing the class distribution. Although the gold standard is normally manually constructed, we were in this case able to construct it automatically due to the fact that a large percentage of posts conform to the Atom Syndication Format.[6] This allowed us to establish the post category automatically by using the *"<category>"* and *"</category>"* tags.

We used this tag plus a counter number corresponding to the consecutive occurrence of a document in the same category to generate a unique identifier i.e. if a document contains the tag *"<category>Politics</category>"* and it is the first document appearing with *"Politics"* tag, the corresponding identifier was set as *"politics1",* the next document with tag *"Politics"* would be *"politics2"*and so on – all these identifiers were placed in the same category called *"Politics"*. If another document appeared with different tag, for instance, *"<category>Sports</category>"* this process is applied to the category *("Sports")* and the identifier was set as *"sports1"* for the first document in *"Sports"* category. In Figure 3 we can see an example of the gold standard constructed for these experiments.

politics1, politics2, politics3, politics4, politics5, politics6, politics7 … politics499, politics500
sports1, sports2, sports3, sports4, sports5, sports6, sports7, sports8, … sports499, sports500
music1, music2, music3, music4, music5, music6, music7, music8, …, music499, music500
sports1, sports2, sports3, sports4, sports5, sports6, sports7, sports8, … sports499, sports500

Fig. 3. Gold standard for English subset.

[6] http://www.atomenabled.org/developers/syndication/

3 Improving the Cluster Quality

In order to test the effectiveness of our approach, we apply an enriching methodology named self-term expansion methodology (S-TEM) in order to improve the quality of the corpus, from a clustering perspective. Then we perform clustering on the enriched corpus and compare the results with the clusters obtained on the original corpus.

3.1 Description of the Clustering Algorithm

In this case we employed the K-means algorithm [14], as it is one of the most popular iterative clustering algorithms. It requires the number of clusters *k* to be fixed a-priori. In general terms, the idea is to choose *k* different centroids. As different locations will produce different results, a useful heuristic 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.

The algorithm steps are as follows:

1. Place K points randomly to represent initial group centroids.
2. Assign each object of the data set to a group that has the closest centroid.
3. When all objects have been assigned to a group, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move.

3.2 The Self-term Expansion Methodology (S-TEM)

The self-term expansion methodology [17] comprises a twofold process: i) the self-term enriching technique (which is a process of replacing terms with a set of co-related terms, and ii) the term selection technique (a process of identifying the relevant features.

The idea behind Term Expansion is not new; it has been studied in previous works such as [21] 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 [3], in which the authors use Wordnet [8] 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*) [15] 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 in the same document; $fr(y)$ and $fr(x)$ are the frequency of y and x respectively and N is a normalization factor equal to the total number of words in the vocabulary.

The self-term expansion technique is defined formally in [18] 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.

To perform term selection, we have used the Document Frequency (DF) technique [10]. This technique has been shown to give good results in [17] and is not computationally expensive. This assigns the value DF(t) to each term t, where DF(t) is the number of posts in a collection where t occurs. It generates a sorted list in descending order and the top terms are selected (in our experiments, we vary the percentage of terms selected in order to empirically establish the optimum value). The assumption is that low frequency terms will rarely appear in other documents and therefore they will not have significance on the prediction of the class of a document. This assumption is valid for all textual datasets, including informal language register data due to fact that rare terms have a low influence in the category definition and are not statistically important for each category. In other words the weight of those terms will be very low, as calculated by the Document Frequency technique.

4 Experimental Results

In this section the performance of the clustering algorithm 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 to informal language register data such as low frequency terms, short vocabulary size and vocabulary overlapping of some domains.

We apply S-TEM in order to replace terms in the corpus with the features discussed in this work by a list of co-related terms; this list may be obtained by general external knowledge resources. However, due to the topic specificity occurring within the informal language register, 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 obtained results on clustering informal language register data with and without S-TEM 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, the FMeasure [23], which is defined as follows: given a set of clusters $C=\{C_1,\ldots,C_{|C|}\}$ and a set of classes $C^*=\{C_1^*,\ldots,C^*_{|C^*|}\}$, the FMeasure between a cluster C_i and a class C_j^* is given in Eq(2).

$$FMeasure(C_i,C_j^*)=\frac{2*precision(C_i,C_j^*)*recall(C_i,C_j^*)}{precision(C_i,C_j^*)+recall(C_i,C_j^*)} \quad (2)$$

The gobal performance of a clustering method is computed using FMeasure values, the cardinality of the set of clusters obtained, and normalizing by the total number of documents $|D|$ in the collection. The obtained result is the FMeasure and it is shown in Eq(3).

$$FMeasure=\sum_{1\le i\le}\frac{|C_i|}{|D|}\max_{1\le j\le|C^*|}FMeasure(C_i,C_j^*) \quad (3)$$

The clustering results of applying S-TEM to our corpus are shown in Figures 4 and 5. Different vocabulary percentages of the enriched corpus were selected by the term selection technique (from 10% to 90%).

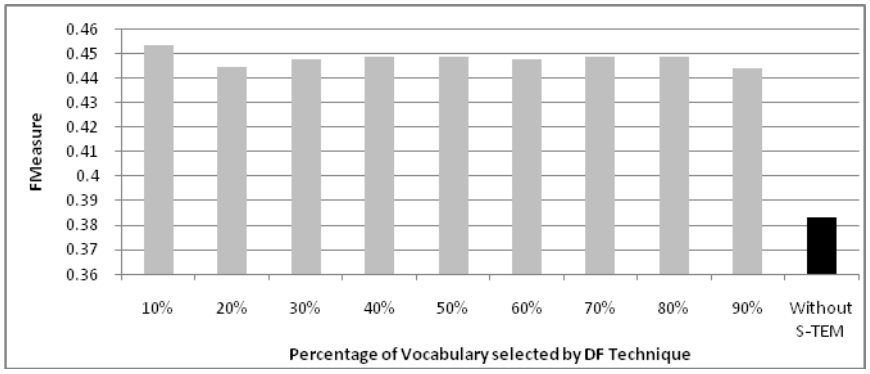

Fig. 4. Clustering results after applying S-TEM to English Subset.

In general, the best clustering results (Figure 4) were obtained by selecting just 10% of the total number of terms belonging to the enriched corpus vocabulary (in English) and using the value of k=4 in the K-means algorithm. This coincides with the number of categories in the gold standard the baseline is 0.383 (Fmeaure) meanwhile the best result obtained by the self-term expansion methodology is 0.453.

In order to be more objective with our methodology, and to show the language independence of the technique (the preprocessing process is not part of our approach), we repeated the experiments, but using an equivalent subset of the corpus, written in the Spanish language. We wish to show that the methodology does not rely on the syntactical structures of any specific language. In other words, it deals with correlated terms and it highlights the intrinsic properties of the same corpus independent of the

language and finally, it selects the most important features in order to improve the accuracy of the clustering algorithms.

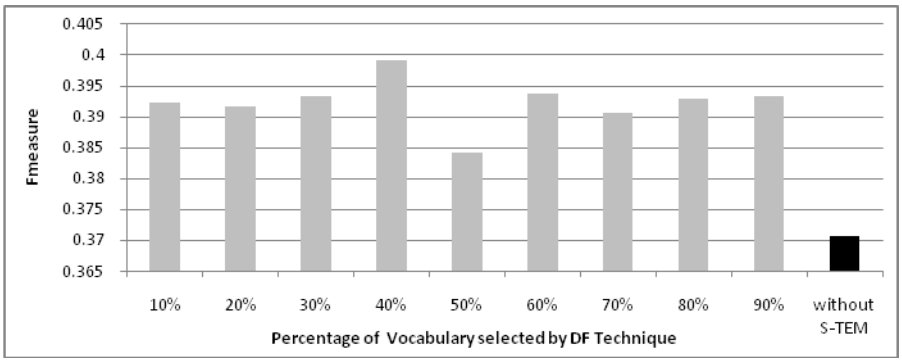

Fig. 5. Clustering results after applying S-TEM to Spanish Subset.

Figure 5 shows the clustering results after applying S-TEM to the Spanish subset. Again the impact of the S-TEM produces improvement in the quality of results, the best FMeasure (0.399) is obtained with 40% of the vocabulary of the enriched corpus, with $k = 4$ (again matching the number of categories in the gold standard). In contrast, the best value obtained by K-means without using our methodology is 0.371. This improvement is due to the addition of the co-related terms to the dataset, although this increases the amount of noise in addition to meaningful information value of the information added to the corpus is considerably higher.

4.1 Pointwise Mutual Information Threshold

We use the Pointwise Mutual Information to provide a value of the correlations between two words. *PMI* calculates the ratio between the number of times that both terms occur together, and the product of the number of times that each word occurs alone. The level of this relationship must be empirically adjusted for each task. We found that a value of 2 or higher to be the best threshold for *PMI* and using bigrams of frequency greater than or equal to 2 for both corpora (English and Spanish). These thresholds were established empirically by analyzing the performance of clustering algorithms with different samples of the datasets. In other experiments, short text of a more structured [17] nature were considered. In these cases, we observed that thresholds for $PMI = 7$ and bigrams of frequency greater than or equal to 4 produced optimum results. However, in our case correlated terms are rarely found inside those documents and we needed to decrement the threshold in order to permit consideration of a longer number of bigrams or correlated terms.

5 Conclusions and Further Work

The process of clustering informal language register data, such as is found in blogs other Web 2.0 sources, is a highly challenging task: texts often exhibit "undesirable" characteristics from a clustering viewpoint, making it difficult for clustering algorithms to achieve good results. Moreover, the specific nature of many of these texts makes it difficult to employ appropriate external linguistic resources.

In this paper, we have described an approach that is not dependent on any external linguistic resources. Moreover we have shown good improvement in clustering informal language register data when our methodology is applied. The self-term expansion methodology uses the corpus itself to improve the quality of the corpus with respect to the clustering task. In addition we have shown that the methodology is language independent, as it does not rely on any syntactical patterns of a specific language, requiring only a preprocessed form of the corpus. Rather, it depends on a list of correlated terms that can contain intrinsic information of each category. The experiments upon which we report in this paper demonstrate that the technique is not dependent on a specific language, as we have obtained improvements when applied to both the English and Spanish subsets of the corpus (although the results achieved with the Spanish subset were not as strong as with the English subset, the still represented an improvement over the baseline).

In future work we plan to use other co-relation measures like those described in [6], in which linguistic properties are used to rank the terms and [16], which introduces a statistical method that score adjoining terms. These will be applied in order to find better relationships between terms, in order to generate the co-related list in the term expansion process, with the goal of providing information to clustering algorithms which is beyond the lexical level. Furthermore, we are investigating the application of the methodology using other clustering techniques such as K-star, in order to demonstrate that the improvements we report are independent of the clustering algorithm.

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 TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 research project.

6 References

1 Alexandrov, M., Gelbukh, A., Rosso, P., An approach to Clustering Abstracts. In: "Natural Lang. Processing and Inform. Systems", LNCS, 3185, pp 85-84 Springer (2005)
2 Balahur, A., Montoyo, A., Multilingual Feature-Driven Opinion Extraction and Summarization from Customer Reviews. Proceedings of NLDB 2008. LNCS, pp 345-346 Springer (2008)

3 Banerjee, S., Pedersen, T., An adapted Lesk algorithm for word sense disambiguation using WordNet. In Proc. of the CICLing 2002 Conference, LNCS, 3878, pp 136–145 Springer (2002)
4 Chesley, P., Vincent, B., Xu, L., Srihari, R. K., Using verbs and adjectives to automatically classify blog sentiment. In Proceedings of AAAI-CAAW-06, pp 27-29, the Spring Symposia (2006)
5 Choi, Y., Breck, E., Cardie, C., Joint Extraction of Entities and Relations for Opinion RecognitionEmpirical Methods in Natural Language Processing (EMNLP) (2006)
6 Daille, B., Qualitative terminology extraction. In Bourigault Jacquemin D., l'homme M.-C., (eds.), Recent Advances in Computational Terminology, volume 2 of Natural Language Processing, pp 149-166 John Benjamins (2001)
7 Devitt, A., Ahmad, K., Sentiment Polarity Identification in Financial News: A Cohesion-based Approach Annual Meeting Association for Computational Linguistics (2007)
8 Fellbaum, C., WordNet: An Electronic Lexical Database. MIT Press (1998)
9 Grefenstette, G., Explorations in Automatic Thesaurus Discovery. Kluwer Academic (1994)
10 Jones, S. K., A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, volume 28 number 1 1972 pp. 11-21 University Press (1972)
11 Koppel, M., Shtrimberg, I., Good News or Bad News? Let the Market Decide AAAI Spring Symposium on Exploring Attitude and Affect in Text (2004)
12 Lawless, S., Hederman, L., Wade, V., OCCS: Enabling the Dynamic Discovery, Harvesting and Delivery of Educational Content from Open Corpus Sources, 8th IEEE International Conference on Advanced Learning Technologies, ICALT 2008: 676-678 (2008)
13 Li, B., Xu, S., Zhang, J., Enhancing Clustering Blog Documents by Utilizing Author/Reader Comments, ACM Southeast Regional Conference, pp 94-99 (2007)
14 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. Berkeley, University of California Press (1967)
15 Manning, D. C., Schütze, H., Foundations of statistical natural language processing. MIT Press (1999)
16 Nakagawa, H., Mori, T., A Simple but Powerful Automatic Term Extraction Method, International Conference on Computational Linguistics, COLING-02 on COMPUTERM 2002: second international workshop on computational terminology - volume 14 (2002)
17 Pinto, D., On Clustering and Evaluation of Narrow Domain Short-Text Corpora, PhD dissertation, Universidad Politécnica de Valencia, Spain (2008)
18 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)
19 Porter, M. F., An algorithm for suffix stripping, Readings in information retrieval, Morgan Kaufmann Publishers Inc., San Francisco, CA (1997)
20 Qamra, A., Tseng, B., Chang, E. Y., Mining Blog Stories Using Community-Based and Temporal Clustering, Conference on Information and Knowledge Management, pp 58-67 Arlington, Virginia, USA (2006)
21 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 (1993)
22 Spanish Stemming Algorithm, http://snowball.tartarus.org/algorithms/spanish/stemmer.html (2009)
23 Van Rijsbergen, C. J., Information Retireval. Butterworths, London (1979)