ISSN 1870-4069

An Effect of Term Selection and Expansion for Classifying Short Documents

Christian Sánchez-Sánchez, Héctor Jiménez-Salazar

Universidad Autónoma Metropolitana Unidad Cuajimalpa,
Departamento de Tecnologías de la Información,
División de Ciencias de la Comunicación y Diseño, Mexico

{csanchez, hjimenez}@correo.cua.uam.mx

Abstract. Many web sites(blogs) over the Internet provide the users the possibility of sharing information like: opinions, news, even their profiles. The peculiarity of this information is that usually the description contains few words. Currently exist a great interest in developing tools that help to process this information in order to organize or categorize it, for helping decision making. Due the importance of this task, in this paper it is explored, through a set of experiments the effect of simple expansion and term selection over two Data Sets. It is applied the Absolute Term Frequency (ATF) term selection technique over this kind of documents, and it is showed that using a percentage of the terms, to represent the information, the classification result could be improved. At the end of the paper it is showed the classification phase where the document expansion could improve the number of classified instances.

Keywords. Term selection, document expansion, document categorization.

1 Introduction

Many Websites(blogs) over the Internet provide the users the possibility of sharing information like: opinions, news, even their profiles. The peculiarity of this information is that usually the description contains few words. For example Twitter allows to write messages at the most 140 characters.

Currently exist a great interest in developing tools that help to process Web Information in order to organize(for instance sentiment analysis) or categorize it(for example topic detection), for helping decision making. Regarding user profiles in some proposals authors try to identify leadership characteristics or classify them according the activity they do. As an example, In 2014 the Reputation Laboratory (REPLAB) [1] it was proposed a task were the objective was to categorize Twitter user profiles according to the domain activities the users do. The categories were: editors, public institutions and so on.

In some proposals for solving those problems supervised classification algorithms have been used. The common phases for classification process are:

document indexing, vector dimension reduction, training, classification and evaluation [5]. Term or feature selection could help to: improve accuracy, delete redundancies and reduce computational cost. On the other hand, Term selection also has been used for query expansion in Relevance Feedback for Information Retrieval.

Nevertheless, regarding to the information contained in Blogs (seen as a collection of short documents) some questions arise: 1) *¿Can Term Selection benefit the categorization of this kind of information?*, 2) if this is so *¿Which term selection technique could work better to improve the classification outcome?*, 3) *¿Can Document expansion improve the classification results?*, 4) *¿What could be more convenient expanding short documents or selecting terms for improving classification results?*

In the interest of finding an answer to these questions, in this paper are employed and tested two different techniques for selecting terms, before classification: Information Gain (IG) and term selection based on taking percentage of term, sorted by Absolute Term Frequency. For expanding documents was used the technique of adding synonyms of all the terms contained in the Data Set.That way, through a set of experiments, it is showed the comparison of the results obtained using the term selection techniques and document expansion. Evaluating the classification outcome (Precision, Recall and F-Measure) using SMO algorithm [11] over two Data Sets (REPLAB14 and 20 News Groups).

It is important to mention that is not the focus of the designed experiments to improve the results of other approaches, but it is to try to identify the effect of term selection and document expansion over short documents during classification process.

The rest of the paper is organized as follow: in next section related work and some techniques for term selection are exhibited, section 3 describes the used Data Sets, section 4 define the proposed experiments, in section 5 the results of the experiments are shown, meanwhile last section present conclusions and future work.

2 Related Work

One of the term selection techniques, usually used for Text Categorization, is Document Frequency (DF). The DF results could be compared with classic techniques like X^2 or Information Gain (IG) [16].

While proposals like the reported by Joachims et al. [3] argues that term selection could weaken the efficiency of Classifiers like Support Vector Machines, also proposals for clustering and classification of Twitter Information have showed that term selection improved the results of those tasks. Such is the case of the work presented by Sánchez-Sánchez et al. [13] where tweets are clustered according to their topic, or the proposed by Villatoro et al. [14] in which Twitter author profiles are classified and ranked, both cases they use DF term selection to improve the results. Similarly, Pinto et al. [10] acceptable

results in the clustering of scientific texts (abstracts) using the Transition Point Technique.

Li et al. [5] proposed to obtain the discriminability and coverage of terms in order to select them, using a combination of measures like DF, a probability ratio and the Average Vector Length. The last measure because it is believed that the poor accuracy at a low dimensionality is imputed to the small average vector length of the documents. They showed that this proposal improved the results gotten using X^2 in two different data sets.

Similarly, Peters et al. [9] presented a uncertainty-based mechanism to discriminate the noisy terms and then select the rest of the terms. Here it is showed how to calculate the uncertainty according to a relative frequency of terms and DF. Such that the model calculates value-uncertainty tuples with the purpose of evaluate the quality of information through a k factor (the value mean divided by uncertainty mean). Small values of k, according to a given value Q, represent noisy terms. It is shown a comparison, against other methods, where it were gotten competitive results over three Data Sets.

The work generated by Lam-Adesina [4] term selection was used in order to tackle the Relevance Feedback IR feature. In Its proposal the first results gotten (using a query) are summarized, The summarizing is done using employing Lunh's keywords clustering [7] with and without considering the query terms. Then the probabilistic model BM25 [12] is applied in order to weight terms and join the heavier terms to a new query.

On the other hand, it is proposed a term selection mechanism based in calculating a set of features: term distribution, query term co-occurrence, pair query term co-occurrence, weighted term proximity, query and expansion DF. Finally, each result is classified as "good" or "bad" to obtain a model that helps the selection.

3 The DataSets

Two collections were used: Twitter Profile and 20 news groups. The first one is composed by a set of user's profiles divides in Training and Test. This data set was generated for been used in REPLAB competition in 2014 (REPLAB14). The data set was designed for solving the task of categorization of users according to a certain domain of activities, classifying users as: publishers, public institutions, athletes, etc.

Below it is given more information about the collections:

3.1 REPLAB14 Training and Test DataSets

Training collection has 10 user's profile categories: celebrity, company, employee, investor, journalist, ngo, professional, public institutions, sportmen, undecidable.

Each category is formed by the next number of documents: celebrity (61 profiles), company (145 profiles), employee (4 profiles), investor (3 profiles),

journalist (466 profiles), ngo (102 profiles), professional (594 profiles), public institutions (40 profiles), sportmen (57 profiles), undecidable (1027 profiles).

Test Data Set has the same number of categories that training, but each category has more profiles. Next more details are given: celebrity (208 profiles), company (222 profiles), employee (14 profiles), investor (7 profiles), journalist (992 profiles), ngo (233 profiles), professional (1543 profiles), public institutions (90 profiles), sportmen (208 profiles), undecidable (1412 profiles).

3.2 20 News Groups

The data set is organized into 20 different newsgroups (corresponding to a different topic). The characteristic of this collection is that some of the topics are very closely related to each other, while others are not. The newsgroups are: comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.-mac.hardware , comp.windows.x, rec.autos, rec.motorcycles, rec.sport.baseball, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, misc.forsale, talk.-politics.misc, talk.politics.guns, talk.politics.mideast, talk.religion.misc, soc.religion.christian and alt.atheism. The version of the data set used has 18846 documents sorted by date(divided into Training and Test) ; duplicates and headers were removed.

4 Experiment Configuration

Some considerations were taken into consideration for the experiments design, these are explained below.

Currently, Support Vector Machines (SVM) [2] had become a learning algorithm very popular for Test Categorization Task [6], due it's consistent execution and capacity for handling big dimension space of inputs. That is why for the proposed experiments was decided to utilize it applying Sequential Minimal Optimization (SMO), using Weka [15]. For tackling multi-class classification was used pairwise classification and the predicted probabilities are coupled using Hastie and Tibshirani's pairwise coupling method in WEKA.

In order to standardize the information and eliminate some elements that may be irrelevant to the classification, the following pre-processing was performed: all text was transformed to lowercase, eliminating URL's, deleting all punctuation marks, removing words and truncating with Porter algorithm.

For representing the documents was used a Boolean weighting model, whereby the presence or absence of the term in the document is indicated. So that with the terms reduction or expansion, the dimension of the vectors changes.

Concerning to answer some of the questions previously stated, the following experiments described in next section were proposed.

4.1 Experiment 1: Classifying Data Sets without Term Selection or Expansion

In the interest to have a baseline, and to know if the reduction or expansion of terms could benefit or affect the classification, as it was stated in Question 1 (which also is immersed in all experiments). The classifier was trained (using cross-validation) to get the models for each of the collections.

That is, for the first five experiments was used REPLAB14 Data Set, the collection was divided into training and test data (5 times, One for each experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. Subsequently,**2a)** it was done the same using 20 News Groups Data Set.

4.2 Experiment 2: Classifying Expanded Documents in Data Sets

In this experiment **2a)** each of the documents contained in the REPLAB2014 collection was expanded, by adding to each document synonyms of the terms that comprise it. In order to get the synonyms Wordnet [8] resource was used. Similarly to the previous experiment 5 experiments were performed, the collection was divided into training and test data (5 times, One for each experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. Subsequently,**2b)** it was done the same using 20 News Groups Data Set.

The results gotten in Experiment **1a** con **2a** were compared with **1b** and **2b** respectively. This in order to help answer questions 3 and 4.

4.3 Experiment 3: Classifying Data Sets formed by Reduced Documents, by Term Selection using Information Gain

The third experiment aimed to help answering questions 2 and 4. In the first part, **3a)** Information Gain method (GI) was used to select a subset of terms, with these terms each document was represented. The representation is based on leaving, inside the Document, only the terms found in the subset obtained. Having done this, five experiments were performed, based on data collection divided into Training and Tests (5 times, one per experiment) The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. Subsequently,**3b)** it was done the same using 20 News Groups Data Set. In **3c)** the documents were expanded and using GI some terms are selected to represent the documents. Finally, **3d)** it was done the same using 20 News Group collection.

4.4 Experiment 4: Classifying Data Sets formed by Reduced Documents, by Term Selection using Absolute Term Frequency(ATF)

Similarly to the previous experiment trying to find answers to questions 2 and 4. **4a)** A list is formed with tuples term-ATF, ATF is the number that the each

term t_k appears in the whole Data Set. Where d_i represents a document, T the Data Set and t_k each term:

$$ATFt_k = \sum_{d_i \in T} t_k. \tag{1}$$

Once all frequencies were obtained the terms were ordered from highest to lowest ATF. Then, it were taken an amount of terms from 10% to 90%. Subsequently, all the documents inside the Data Set were represented with each percentage of terms. After that, five experiments were performed, based on data collection divided into Training and Tests (5 times, one per experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. **4b)** it was done the same using 20 News Groups Data Set.

For each classification obtained(using the percentages of terms to represent documents), the results were compared in order to identify the best representation, that which improved precision. In some cases finest results were searched, through exploring nearest percentages of terms, taking one percent of terms more (of the best) each time.

Results from experiments **3a** and **4a**, and **3b** and **4b** were compared respectively, in order to identify the representation that improves precision.

4.5 Experiment 5: Classifying reduced Documents after Expansion

This experiment was designed looking for finding an answer to the question 3. Using REPLAB14 Data Set it was sought to test if **5a)** the winning term selection technique(reduction) worked after Document expansion. That's why also, five experiments were performed, based on data collection divided into Training and Tests (5 times, one per experiment). The classification was evaluated and at the end the results were averaged. As results were obtained *Precision*, *Recall* and *F-Measure*. **5b)** it was done the same using 20 News Groups Data Set.

4.6 Experiment 6: Training the Classifier with the Best Representations for the Data Sets and Testing

With regard to integrate and use results and answering questions 1-4.In this experiment it was sought to use the best representation of the Training part in order to generate the model and evaluate the results of classifying the test part, for both collections: **6a1)** REPLAB14 and **6a2)** 20 News Group.

For making this possible, the Test part of both Data Sets must be represented with the same terms of the winner reduction (selection) technique. Meanwhile, in the last experiment the documents inside the Test part of each Data Sets:**6b1)**REPLAB14 and **6b2)**20 News Groups are expanded and then represented with the winner reduction technique.

5 Experiment Results

Next the results of each experiment are shown, as well as the comparisons among: term selection techniques, term selection vs expansion, and term selection vs no selection.

Experiment 1

Those results were the baseline to compare, wondering if term selection improves classification. After pre-processing and indexing Data Sets, the classification result (P=Precision, R= Recall, F= F-Measure) was:

1a)
P=0.299, R=0.205, F=0.216

1b) While using 20 News Groups result was:
P=0.767, R=0.751, F=0.756

Experiment 2

2a) After expanding documents, including synonyms of each term inside each document, and then classifying REPLAB14, the outcome was:
P=0.241, R=0.192, F=0.201.

If this result is compared to the previous **1a** it can be observed that the classification was not improved after expansion. Actually, F-Measure was lower, $0.201 < 0.216$.

2b) Afterwards expanding 20 News Groups, results were:
P=0.717, R=0.746, F=0.731 Comparing F-Measure to the previous **1b** it can be noted that the classification neither was improved, , $0.731 < 0.756$.

Experiment 3

3a) After applying Information Gain technique, for selecting terms in REPLAB14, the results were:
P=0.304, R=0.201, F=0.213 Comparing results to the reported in **1a** neither was an improvement in F-Measure, though the precision was higher.

3b) Applying GI for term selection to represent 20 News Groups Documents results were:
P=0.769, R=0.75, F=0.759.In this case there was improvement in classification results compared to**1b**.

3c) If GI is applied after expanding Documents the result was:
P=0.223, R=0.19, F=0.197 It can be observed that the classification result is worst than the reported in **1a**.

3d) If a similar process **3c** is applied to 20 News Groups Data Set the result is not favorable:
P=0.761, R=0.75, F=0.755

Experiment 4

4a) After calculating the Absolute Term Frequency for all terms, the terms are listed from lowest to highest frequency. The results of taking a percentage of those terms and represent the collection REPLAB14 and classify it is showed in the next table (see Table.1). As it can be seen the classification outcome, compared with the result reported in **1a**, is highest if it is taken from 60% to

70% where it reach the peak (P=0.317, R=0.224, F=0.242) and then it start decreasing.

Table 1. Classification results gotten employing ATF term selection, over REPLAB14.

Percentage	Precision	Recall	F-measure
10%	0.249	0.218	0.227
20%	0.268	0.219	0.232
30%	0.259	0.215	0.226
40%	0.28	0.223	0.237
50%	0.297	0.221	0.236
60%	0.271	0.208	0.219
70%	**0.317**	**0.224**	**0.242**
80%	0.286	0.212	0.224
90%	0.286	0.205	0.217
100%	0.299	0.205	0.216

4b) Applying the same term selection technique (ATF), in order to classify 20 News Groups the results are depicted in Table 2.

Table 2. Classification results gotten employing ATF term selection, over 20 News Groups.

Percentage	Precision	Recall	F-measure
10%	0.737	0.728	0.731
20%	0.747	0.736	0.74
30%	0.748	0.736	0.74
40%	0.75	0.737	0.742
50%	0.75	0.739	0.743
60%	0.754	0.743	0.747
70%	0.758	0.743	0.748
80%	0.759	0.744	0.748
90%	**0.771**	**0.754**	**0.759**
100%	0.767	0.751	0.756

It can be detected that the classification outcome is improved, compared with the reported in **1b**, if the percentage of selected terms is 90% where it is the maximum (P=0.771,R=0.754,F=0.759). The term selection technique applied in both Data Sets helped to improve the results.

Experiment 5

5a) If document expansion is applied before ATF term selection, with the purpose of classifying REPLAB14, the next results were gotten (See Table.3).

Comparing the results reported in **1a**, it can be seen that representing documents with the 80% of the terms the classification is better (P=0.242, R=0.2, F=0.21). In spite of the documents where previously expanded the result was

the opposite than the reported in **3c**. Nevertheless, the result was lower than the reported in **4a**.

Table 3. Classification results gotten employing ATF term selection after document expansion, over REPLAB14.

Percentage	Precision	Recall	F-measure
10%	0.179	0.174	0.174
20%	0.181	0.177	0.177
30%	0.208	0.199	0.201
40%	0.222	0.202	0.208
50%	0.222	0.202	0.208
60%	0.212	0.192	0.198
70%	0.227	0.194	0.201
80%	**0.242**	**0.2**	**0.21**
90%	0.232	0.19	0.198
100%	0.241	0.192	0.201

5b) If it is done the same process, described in **5a**, over 20 News Groups the maximum outcome was obtained selecting 77% of the terms, (P=0.633,R=0.629, F=0.630). Nonetheless, it is not a better result that the reported in **4b**.

It is important to say that for those experiments the ATF term selection worked better than IG, and it has benefited the classification results of such documents.

Experiment 6

6a1) Due that selecting the 70% of the terms sorted by ATF, and representing the REPLAB14 documents, in order to classify them, gave the best result. Then all the documents of the Test Data Set part were represented with those terms. It was used the model gotten using the REPLAB14 training part and the results of classifying the test was:
P=0.333 R=0.141 F=0.125

It is important to mention that selecting the 70% of the terms 4619 documents of a total of 4929, contained in the Test Part. And the number of correctly classified instances was 1638 (35.5 %)

6a2) The result of applying the previous process over 20 News Groups was:
P=0.769 R=0.757 F=0.762

6b1) Nevertheless if the REPLAB14 Test documents are represented using the 70% of the terms after Document expansion, using the same model used in **6a1** for classification. The results were:
P=0.260 R=0.174 F=0.172

Although the precision was lower indeed more documents could be represented, compared to **6a1**, 4749 documents of a total of 4929. Where 1873 Documents were classified correctly (39.5%).

In this case it can be said that the expansion in the test documents helped to classify correctly more documents.

6b2) The result of applying the previous process over 20 News Groups was: P=0.748 R=0.793 F=0.0.769

6 Conclusions and Future Work

As a conclusion it could be argued that employing ATF Term Selection and GI benefited classification results. This using SMO Algorithm and a Boolean Weight Representation, of the vector members for representing the documents. The ATF Term selection technique was better although other terms where included in the documents (by expansion).

In this case ATF term selection was better, selecting from 80% to 90% the classification accuracy is better than using all the terms. This was a constant in both Data sets.

ATF algorithm took the terms according to their popularity (frequency between documents and frequently in the document, jointly) and therefore did not discriminate some terms that other techniques may penalize or dismiss easily. This could be deduced reviewing the REPLAB14 Data Set but it is necessary to analyse other Data Sets.

Document expansion (adding the synonyms of all the term into the document) showed that can help to improve classification results if and only if it is applied in the Test part, similarly that it is done in Relevance Feedback for Information Retrieval.

As future work it is planned to: *a)* Perform tests of statistical significance for the results and subsequently, *b)* design experiments with; other collections of similar short documents, other types of representations and other classifiers. *c)* make comparisons to other term selection techniques like DF or Transition Point.

References

1. Amigó, E., Carrillo-de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., Rijke, M., Spina, D.: Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings, chap. Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management, pp. 307–322. Springer International Publishing, Cham (2014), http://dx.doi.org/10.1007/978-3-319-11382-1_24
2. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995), http://dx.doi.org/10.1023/A:1022627411411
3. Joachims, T.: Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings, chap. Text categorization with Support Vector Machines: Learning with many relevant features, pp. 137–142. Springer Berlin Heidelberg, Berlin, Heidelberg (1998), http://dx.doi.org/10.1007/BFb0026683

4. Lam-Adesina, A.M., Jones, G.J.F.: Applying summarization techniques for term selection in relevance feedback. In: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1–9. SIGIR '01, ACM, New York, NY, USA (2001), http://doi.acm.org/10.1145/383952.383953
5. Li, J., Sun, M.: Scalable term selection for text categorization. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL). pp. 774–782. Association for Computational Linguistics, Prague, Czech Republic (June 2007), http://www.aclweb.org/anthology/D/D07/D07-1081
6. Liu, Y., Loh, H.T., Kamal, Y.T., Tor, S.B.: Natural Language Processing and Text Mining, chap. Handling of Imbalanced Data in Text Classification: Category-Based Term Weights, pp. 171–192. Springer London, London (2007), http://dx.doi.org/10.1007/978-1-84628-754-1_10
7. Luhn, H.P.: The automatic creation of literature abstracts. IBM J. Res. Dev. 2(2), 159–165 (Apr 1958), http://dx.doi.org/10.1147/rd.22.0159
8. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (Nov 1995), http://doi.acm.org/10.1145/219717.219748
9. Peters, C., Koster, C.H.A.: Advances in Information Retrieval: 24th BCS-IRSG European Colloquium on IR Research Glasgow, UK, March 25–27, 2002 Proceedings, chap. Uncertainty-Based Noise Reduction and Term Selection in Text Categorization, pp. 248–267. Springer Berlin Heidelberg, Berlin, Heidelberg (2002), http://dx.doi.org/10.1007/3-540-45886-7_17
10. Pinto, D., Jiménez-Salazar, H., Rosso, P.: Clustering Abstracts of Scientific Texts Using the Transition Point Technique, pp. 536–546. Springer Berlin Heidelberg, Berlin, Heidelberg (2006), http://dx.doi.org/10.1007/11671299_55
11. Platt, J.C.: Sequential minimal optimization: A fast algorithm for training support vector machines. Tech. rep., Advances in kernel methods - Support vector learning (1998)
12. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. 3(4), 333–389 (Apr 2009), http://dx.doi.org/10.1561/1500000019
13. Sánchez-Sánchez, C., Jiménez-Salazar, H., Luna-Ramírez, W.A.: Uamclyr at replab2013: Monitoring task. In: CLEF (Working Notes) (2013)
14. Villatoro-Tello, E., Ramírez-de-la Rosa, G., Sánchez-Sánchez, C., Jiménez-Salazar, H., Luna-Ramírez, W.A., Rodríguez-Lucatero, C.: Uamclyr at replab 2014: Author profiling task. In: CLEF (Working Notes). pp. 1547–1558 (2014)
15. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
16. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proceedings of the Fourteenth International Conference on Machine Learning. pp. 412–420. ICML '97, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1997), http://dl.acm.org/citation.cfm?id=645526.657137