Enhancement of DTP Feature Selection Method for Text Categorization*

Edgar Moyotl-Hernández, Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación,
B. Universidad Autónoma de Puebla,
emoyotl@mail.cs.buap.mx, hjimenez@fcfm.buap.mx

Abstract. This paper studies the structure of vectors obtained by using term selection methods in high-dimensional text collection. We found that the *distance to transition point* (DTP) method omits commonly occurring terms, which are poor discriminators between documents, but which convey important information about a collection. Experimental results obtained on the Reuters-21578 collection with the k-NN classifier show that feature selection by DTP combined with common terms outperforms slightly simple *document frequency*.

1 Introduction

The goal of *text categorization* (TC) is to classify documents into a set of predefined categories. In TC each document is usually represented as a vector of terms in a multidimensional space, in which each dimension in the space corresponds to a term. Typically even a moderately sized collection of text has tens or hundreds of thousands of terms. Hence, the document vectors are high-dimensional. However, most documents contain fewer terms, 1-5% or less, in comparison to the total number of terms in the entire text collection. Thus, the document vectors are *sparse* [3].

For reasons of both efficiency and efficacy, *feature selection* (FS) techniques are used when applying machine learning algorithms to text classification. In our previous experiments [6] we found that FS using DTP achieves performance superior to *document frequency*, and comparable to *information gain* and *chi-statistic*; three well known and effective FS techniques [10]. However, the vectors produced by DTP have a "sparse" behavior that is not commonly found in low-dimensional text collections.

In this paper, our first focus is to study the structure of the vectors produced by term selection methods when applied to large document collections. Such structural insight is a key step towards our second focus, which is to explore the relationships between DTP and the problem of the sparseness. We hypothesized that supplementing it with high frequency terms would improve term selection by adding important (and also common) terms; and we report experimental results

* This work was supported by VIEP-BUAP, grant III9-04/ING/G.

obtained on the standard Reuters-21578 benchmark with the k-NN classification algorithm.

The paper is organized as follows. Section 2 compares the sparseness and weighting of the vectors produced from the output of the term selection techniques: *document frequency*, *information gain*, *chi-statistic*, and DTP. Furthermore, section 2 shows that vectors obtained by DTP are sparse and that vectors obtained by combining DTP with *document frequency* are dense. Section 3 presents conclusions and future research.

2 Density and Weighting of Vectors

In this section, we empirically study the structure of the vectors produced by term selection methods. As we will see density of vectors is calculated instead of sparseness.

We used the Reuters-21578 collection which consists of 12,902 news stories classified according to 115 thematic categories. The experiments used the set of the 10 most frequent categories (R10), partitioned (according to the ModApte split) into a training set of 6,490 documents and a test set of 2,545 documents. Term weighting was done using $tfidf_{ij} = tf_{ij} * \log(N/df_i)$, where tf_{ij} is the number of times t_i occurs in a document d_j, df_i is the number of documents in which t_i occurs, and N is the number of documents [7]. Also tf_i is defined as $\sum_j tf_{ij}$. We will refer to four FS methods (see [10] for more details), which can be briefly defined as follows. Document frequency (DF) is the number of documents in which a term t_i occurs (df_i). Chi-statistic (CHI) measures the lack of independence between a term and the category. CHI computed for a term takes the maximum on all categories; CHImax. Information gain (IG) of a term measures the number of bits of information obtained for category prediction by knowing the presence or absence of the term in a document. IGsum for a term represents the expected information gain on categories. As we have said, these three FS methods are effective in the TC task [10]. DTP is based on the proximity to the frequency that splits the terms of a text into low and high frequency terms; this frequency is called the *transition point* (TP). DTP is computed by the distance from frequency (tf_i) of term (t_i) to TP. Given a text, TP is easy to calculate because it only requieres the number of words t_i with $tf_i = 1$ [1][9][5]. We refer to the above technique as the *Inverse DTP* (IDTP) rule. More important terms for the TC task are those producing the lowest DTP scores. IDTP showed a comparable performance [6] to the best term selection techniques, CHI and IG, although this fact depends on the size of text collection.

Table 1 shows density (columns 2-6) and average weighting (columns 7-11) for three percentages[1] of terms selected by DF, IGsum, CHImax, and IDTP. One more FS method was included, IDTP$_{df}$*DF, which will be discussed after Density was calculated as the ratio of the number of nonzero terms in training and test vectors to the total number of selected terms. Zipf's Law implies that

[1] Since FS methods give better performance in the TC task taking 1%, 5% and 10% of highest score [10], we used such percentages in the experiments.

more than 50% of terms have frequency 1, more than 10% have frequency 2, etc. [1]. So the more terms selected, the higher the percentage of low frequency terms. Therefore, the density of vectors with terms given by any FS method decreases as the percentage of terms grows; which can be seen in columns 2 to 5 of table 1. It must be remarked that DF has the highest density and IDTP has the lowest density. This means that selecting terms using IDTP will give us sparse vectors. Besides, because term selection by CHImax or IGsum takes into account categories intended to match the right class in the TC task, they do not depend on weighting (cols. 8 and 9 in table 1); weighting is distributed among categories. A growing tendency of average weighting of vectors is observed in DF and IDTP values (cols. 7 and 10). Since IDTP is based on the importance of terms, from the vector space model point of view, then less frequent terms are more important, i.e. medium frequency terms are the weightiest.

Table 2 summarizes microaveraged F_1 values obtained for the $k-$NN classifier (using $k = 30$) with the evaluated FS techniques for different percentages of terms. Columns 2, 3, 4 and 5 correspond to DF, IGsum, CHImax and IDTP, respectively. DF, IGsum and CHImax values for k-NN are in accordance with the findings of Debole and Sebastiani [2]. The results of IDTP imply we should reinforce term selection with frequent terms. A simple way to attain this purpose is by providing a higher score to frequent terms; for example IDTP*DF as the score. Although the IDTP*DF score is better than IDTP, this score is only comparable to DF at 10%; see columns 2, 5 and 6 of table 2. Thus, the IDTP score was reformulated considering that tf_i represents the intratext frequency, while df_i represents the intertext frequency. We define IDTP_{df} for a term t_i as the inverse distance between df_i and TP_{df}, where TP_{df} is computed as the transition point using df_i instead of tf_i. In order to select important terms, we use IDTP_{df} multiplied by DF, which raises the score of frequent terms. Results of IDTP_{df}*DF are shown in columns 6 and 11 of table 1 and in column 7 of table 2. We see that F_1 for IDTP_{df}*DF is as good as for DF. Also, values for IDTP_{df}*DF show the increase of the density of the selected terms (col. 6 of table 1), and a more stable average weighting (col. 11 of table 1).

3 Conclusions

A feature selection method based on IDTP was proposed. It was motivated by remarks about density and weighting of vectors built with terms near the transition point. This feature selection method multiply IDTP by DF and thus improves on DF, one of the most effective term selection methods in TC tasks.

An advantage of IDTP is the low computational cost compared with top feature selection methods (CHI or IG). However, there are several point pending such as testing with other high-dimensional text collections, applying criteria for selecting terms which take the category into account, and to experiment with differents ways to use TP.

Percent	Density					Weight (avg.)				
of Terms	DF	IGsum	CHImax	IDTP	IDTP$_{df}$*DF	DF	IGsum	CHImax	IDTP	IDTP$_{df}$*DF
1	0.1	0.075	0.056	0.012	0.016	4.2	5.4	5.7	5.1	5.3
5	0.035	0.030	0.021	0.009	0.036	5.5	5.5	5.8	7.2	5.2
10	0.021	0.020	0.017	0.006	0.021	5.8	5.7	5.6	8.0	5.8

Table 1. Term selection *vs* density and average weighting of training and test vectors.

% terms	DF	IGsum	CHImax	IDTP	IDTP*DF	IDTP$_{df}$*DF
1	0.826	0.855	0.855	0.302	0.314	0.392
5	**0.851**	**0.860**	**0.863**	0.499	0.738	0.851
10	0.850	0.853	0.859	**0.545**	**0.845**	**0.853**

Table 2. Microaveraged F_1 for several FS criteria using $k-$NN on R10.

Acknowledgements. We would like to thank James Fidelholtz by useful comments on this work.

References

1. Booth, A.: A Law of Occurrences for Words of Low Frequency, Information and Control, (1967) 10(4) 386-93.
2. Debole, F. and Sebastiani, F.: An Analysis of the Relative Difficulty of Reuters-21578 Subsets, In Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation, Lisbon, PT, pp. 971-974.
3. Dhillon, I. S., Modha, D. S.: Concept Decompositions for Large Sparse Text Data using Clustering. Mach. Learn., Kluwer Academic Publishers, (2001) 42(1-2) 143-175.
4. Galavotti, L., Sebastiani, F., Simi, M.: Experiments on the Use of Feature Selection and Negative Evidence in Automated Text Categorization, Proc. of ECDL-00, 4th European Conference on Research and Advanced Technology for Digital Libraries, (2000) 59-68.
5. Moyotl, E., Jiménez, H.: An Analysis on Frequency of Terms for Text Categorization, Proc. of SEPLN-04, (2004).
6. Moyotl, E., Jiménez, H.: Experiments in Text Categorization using Term Selection by Distance to Transition Point, Proc. of CIC-04, (2004).
7. Salton, G., Wong, A., Yang, C.: A Vector Space Model for Automatic Indexing, Communi-cations of the ACM, (1975) 18(11) 613-620.
8. Sebastiani, F.: Machine Learning in Automated Text Categorization, ACM Computing Surveys, Vol. 34(1), (2002) 1-47.
9. Urbizagástegui-Alvarado, R.: Las posibilidades de la ley de Zipf en la indización automática, Reporte de la Universidad de California Riverside, (1999).
10. Yang, Y., Pedersen, P.: A Comparative Study on Feature Selection in Text Categorization, Proc. of ICML-97, 14th Int. Conf. on Machine Learning, (1997) 412-420.
11. Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods, Proc. of SIGIR-99, 22nd ACM Int. Conf. on Research and Development in Information Retrieval, (1999) 42-49.
12. Zipf, G.K.: Human Behaviour and the Principle of Least Effort, Addison-Wesley, (1949).