Winning some of the document preprocessing challenges in a text mining process

Bruno M. Nogueira[1], Maria F. Moura[2], Merley S. Conrado[1], Rafael G. Rossi[1], Ricardo M. Marcacini[1], Solange O. Rezende[1]

[1]Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação.
Caixa Postal 668, São Carlos, SP, Brasil - 13560-970

[brunomn,merleyc,solange]@icmc.usp.br

[2]Embrapa Informática Agropecuária.
Caixa Postal 6041, Campinas, SP, Brasil - 13083-970

fernanda@cnptia.embrapa.br

Abstract. *Considering the huge growth of the number of documents in the digital universe and the possibility of obtaining some competitive advantage in processing them, this paper describes some of the difficulties of working with text collections. More specifically, it shows some of the challenges on the step considered one of the most important of the Text Mining process - the data preprocessing - focusing on two of its main tasks: attribute generation and selection, considering not only single terms but composed terms too. In order to overcome the challenges imposed by these problems, this paper presents efficient unsupervised solutions. The application of these solutions in three real data sets is presented in order to evaluate them and to show a way to treat the data step by step. Good results were obtained at the end of the whole process.*

1. Introduction

In a context where an increasing amount of textual data is stored by different organizations, the Text Mining (TM) process using computational techniques of knowledge extraction, acts as a transformer agent. Useful knowledge is extracted from this enormous quantity of textual data, being used as a competitive advantage or as a support to decision making. This process can be seen as a particular case of Data Mining which is composed by five steps: problem identification, pre-processing, pattern extraction, pos-processing and knowledge use. These steps can be instanced according to the process goals [11].

Frequently, the pre-processing is dealt as a step of minor importance, or less interesting than the others, due to the lack of technical glamor and the excess of manual tasks. Basically, this step aims at transforming the text collection into a useful form for the learning algorithms, involving tasks as treatment, cleaning and reduction of the data. In this work, two of the main pre-processing difficulties are highlighted: attribute generation and attribute selection. This work attempts to obtain meningful aingle and composite terms (unigrams and bigrams) from the text collection at the attribute generation step. So, with the generated terms, the most representative terms are selected through the application of some attribute selection methods. It is necessary to highlight that previous works, such as [3], [4] and [5], show the application of some of these attribute selection methods

facing only single terms. In the present work, the capability of these methods in selecting representative composite terms is also measured.

This work is inserted in a context that aims to extract a topic taxonomy for a domain which is represented by textual documents, following the methodology proposed by Moura[9]. The methodological basis is a semi-automated Text Mining process that aids the identification of the domain taxonomy within a text collection using a hierarchical clustering algorithm. The main objective of the extracted taxonomy is to help the domain specialists to manage document collections and the information contained in these documents. The domain specialist, with the help of statistical measures, can intervene in some process steps and also edit the generated taxonomy in order to adjust the results to the problem requirements. Thus, generating meaningful single and composed terms and selecting the most representative ones are very important to assure two important aspects for this methodology: terms comprehensibility and representativeness. Since this methodology deals with unlabeled collections, the solutions presented here are unsupervised methods.

In the next section the methodology used in the pre-processing step and its evaluation process are described; followed by the experiments and their results, and finally the conclusions.

2. Pre-processing Methodology

In this section the solutions used to pre-process non-classified text collections of a knowledge domain are described. Due to the context of the topic taxonomy extraction, the pre-processing must assure the quality of the data concerning the comprehensibility and the representativeness. Additionally, all the tasks of the pre-processing step, from the choice of the text collection to the attribute selection, must allow the domain specialist to intervene in the process, if he desires.

According to these requirements, the pre-processing was considered in three tasks: text collection standardization, attribute generation and attribute selection. Before detailing these steps, it is necessary to highlight that the attribute representativeness is not easy to measure due the unlabeled text collection. So, in order to validate the choice of the methods and, consequently, the choice of the attribute set, the process is used over a labeled collection submitted to a classifier, as described in the validation subsection. However, these class labels were not used in the attribute generation and selection steps.

2.1. Text collection standardization

The performed standardization process depends on the goal and how the data must be represented. The available documents are subjectiely analyzed, that is, it is verified if they are representative and not damaged. The available documents are very often in several different formats that are not directly usable, requiring some conversions. This is an ever recurrent step until the text collection inspires confidence.

At the first step, a **document conversion** is carried out by converting all documents to plain text format and discarding those that can not be converted. The remaining texts are submited to a **language separation** process that stores documents in different collections according to the language they were written. After that, a **character standardization** process is applied, removing all unnecessary characters from the documents

- such as accents, punctuation marks, cedillas, numbers, underlines and mathematical symbols - and transforming all remaining characters to lower case.

With the texts in a standard format, a **verification for pre-existent information** has to be done, looking for some pre-existent metadata, such as author and title, and inserting them in the respective document. Finally, an subjective **evaluation** is done, verifying if the final collection is adequate to represent the problem domain. If the database is considered insufficient, somehow it has to be completed. This evaluation can imply in repeating the process over and over again, until the collection is subjectively considered satisfactory to reach the identified goals.

2.2. Attribute generation

With a confident text collection, *a priori* sufficient, it is necessary to generate the attributes, that are the terms of interest. **Terms**, in this work, are used in the same context as for information retrieval: they can be either simple or stemmized words, considered alone or in a phrase combination, and treated as a ngram. For example, a term can be: "decis", "make", "system", "applic","artifici", "decis-make", or "decision", "making", "application", "artificial", "decision-making".

In this step, the PreText software tool [8] was used to identify onegrams and twograms. This tool allows the elimination of the stopwords before the identification of the onegrams, their stemmization and the twogram combinations. The stemming process is based on Porter´s [10] algorithm and adapted to three languages: Portuguese, Spanish and English. All ngrams are generated considering each text in the collection as a bag-of-words. Additionally, for each ngram, the tool calculates its occurrence frequency in the collection and the number of texts in which the ngram is presented.

As the number of generated twograms is huge and most of them have no semantical meaning, some choice tests have to be applied, considering the occurrence position of each onegram in the whole collection. These choices try to identify potential meaningful and representative composed terms among the generated ngrams. For example, in the phrase with no stopwords: "artificial intelligence technique applications have been used decision making systems"; "artifici-intellig" and "decis-make" are potential terms, but "applic-decis", "make-system" and "techniqu-applic" do not add relevant domain information. So, if the test results indicate them as non-relevant, these ngrams have to be discarded.

In order to carry out this selection, the chosen statistical test was the log of the likelihood ratio as it is robust enough to be used over sparse data [7] and its implementation is available at Ngram Statistical Package - NSP [2]. Basically, that test indicates the dependence ratio of each onegram related to the other onegram in the combination, considering their position in the twogram. As the NSP tool provides a scored dependence relation of the twograms as the result report, those ngrams which have the score values greater than 3.84 are chosen to be effectively used from this step on. This assumption is equivalent as to reject the independence test formulating the hypothesis considering a qui-square distribution of one grade of freedom with certain of 95%.

2.3. Attribute selection

Even after a thorough process of cleaning, attribute generation and eliminating the statistically insignificant twograms, the attribute number is still huge. Not all of these terms are present in each document of the collection, resulting in sparse representations of the frequency values. Thus, choosing a good attribute filter in this step implies in selecting better attributes to delimitate the problem domain and, consequently, contributing to improve the performance of the learning algorithms used in the knowledge extraction step. Moreover, the filter has to assure the term representativeness, even if the collection is not labeled, which implies in one more difficulty: how to delimit a representative set of terms.

The most common used filter is the Luhn´s cutoffs [6]. To find these cutoff points, the occurrence frequencies of the attributes are ascending classified and plotted. So, two cutoffs points are chosen close to the tendency curve inflexion points, considering the attributes which have very low or very high frequencies as irrelevant. Despite this, the elimination of low frequency terms is not a common-sense. For example, in the information retrieval area they have been favored, because of the spreadly use of the tf-idf indexer (tf-idf: term frequency - inverse document frequency) [12].

In this step, some filters based on variance representativeness are also evaluated:

1. Term Contribution - TC [5]: the Term Contribution can be defined as how much one specific term contributes to the similarity among documents in a document collection. It can be calculated as show in Eq. 1:

$$TC(t_k) = \sum_{i,j \cap i \neq j} f(t_k, D_i) * f(t_k, D_j) \tag{1}$$

   where, $f(t_k, D_i)$ is the TF-IDF of the *k-th* term of the *i-th* document.
2. Term Variance - TV [4]: this measure is used to calculate the variance of all terms in the collection, giving the highest scores to those terms that do not have low document frequency and have a non-uniform distribution through the collection. It can be expressed as shown in Eq 2:

$$v(t_i) = \sum_{j=1}^{N} \left[f_{ij} - \bar{f}_i \right]^2 \tag{2}$$

   where, f_{ij} is the frequency of the *i-th* term of the *j-th* in document and $\bar{f}_i$ is the mean frequency of the terms in the document collection.
3. Term Variance Quality - TVQ [3]: the Term Variance Quality is very similar to Term Variance and uses the total variance to calculate the quality of a term, as shown in Eq. 3

$$q(t_i) = \sum_{j=1}^{n} f_{ij}^2 - \frac{1}{n} \left[\sum_{j=1}^{n} f_{ij} \right]^2 \tag{3}$$

   where f_{ij} is the frequency of the *i-th* term of the *j-th* document.

Therefore, these filters provide an attribute ranking which implies in a cutoff choice. So, the number of remaining attributes after the Luhn´s cutoffs, called here k, is used to estimate the number of attributes to be considered in the calculated rankings. That is, the k top-ranked attributes will be taken for each calculated ranking.

Besides these four filters, two others based on Document Frequency (DF) of a term were also evaluated. The first one is based on Salton's cutoffs [13], which consider terms whose DF is between 1% and 10% of the total number of documents.

The other filter, proposed by the authors, is an adaptation of Luhn's cutoff idea for DF. In this sense, the ascending ordered histograms of terms' DF are plotted and two cutoff points are chosen next to the inflexion points of the tendency curve. This cutoff will select terms whose DF is neither too small nor too high.

Using these filters, the obtained subsets are evaluated, using the proposal validation process which is explained in Section 2.4.

2.4. Attribute set validation

In order to carry out a non subjective evaluation of the pre-processing results and obtain a validation of the generated attribute set, a labeled document collection is used within a supervised learning process and evaluation. In this way, all the described steps are applied to a labeled text collection, but considering this text collection as non labeled. With the chosen attributes, an attribute-value matrix is constructed, where, each row vector represents a document and each column an attribute; the cells correspond to the occurrence frequency of the attribute in the document; and finally the last column corresponds to the codified label.

So, to validate the results, for each obtained attribute subset containing the generated onegrams and twograms, two classification models are constructed using two widely known classification algorithms: C4.5 decision trees and Support Vector Machines. Both of them were chosen because they can face well sparse domains. Additionally, to estimate the classifiers accuracy rate, the 10-fold cross validation process is used.

3. Experiments and Results

In this section, the experiments carried out to evaluate the six unsupervised attribute selection methods presented on Section 2.3 are shown.

3.1. Text collection pre-processing

Three data sets from different domains and sizes were selected. The first is a collection of articles from the Instituto Fábrica do Milênio (IFM) [1]; that is a Brazilian organization whose actions are focused on the search for manufacturing solutions for the industry needs. This data set is composed of 614 documents in the Portuguese language, divided into 4 classes, with 291 documents in the majority class. The second document collection is the Case Based Reasoning- Inductive Logic Programming - Information Retrieval - Sonification (CBR-ILP-IR-SON) data set[2] composed of 681 documents in the English language, classified according to 4 classes with 276 documents in the majority class. The third data set is the Twenty Newsgroups[1] where 50 documents were randomly selected from each class (newsgroups), totalling 1000 documents.

1. Text collection standardization: initially, damaged or duplicated documents were discarded and, finally, the transformations enumerated in section 2.1 were applied.

[1] http://www.ifm.org.br

[2] http://infoserver.lcad.icmc.usp.br/infovis2/PEx

The CBR-ILP-IR-SON was reduced from 681 to 675 documents, although the Twenty Newsgroups and IFM document collections had not been reduced.

2. Attribute generation: first, the PreText tool was used to remove the stopwords and to apply the stemming process, obtaining all possible onegrams and twograms from each text collection. After that, the NSP tool was used to rank the twograms according to the log of the likelihood ratio dependence test. Finally, the twograms which rankings were greater than 3.84 were taken. In Table 1 the results obtained with the attribute generation are shown. The number of twograms was drastically reduced as expected, because the number of possible combinations is huge and most of them are statiscally non-relevant.

Data sets	Number of grams		
	Onegram	Twograms	Twograms chosen
IFM	34789	606404	26203
CBR-ILP-IR-SON	23155	104461	31238
20 Newsgroups	17410	63489	15707

Table 1. Description of generated attributes

3. Attribute selection: in this step, the six attribute selection algorithms presented in section 2.3 were applied over the three data sets. As previously shown, Luhn-TF, Salton and Luhn-DF methods have suggest fix cutoff points, while TC, TV and TVQ methods do not. For these algorithms that only generate rankings and do not have pre-defined cutoff points, a subset was chosen. This subset contained the k better ranked attributes and k is the same number of attributes selected by the Luhn-TF method. A summary of this attribute selection is shown in Table 2.

Cutoff	IFM	CBR-ILP-IR-SON	20 Newsgroups
	Subset Size	Subset Size	Subset Size
Luhn-TF	16540	10760	4689
Salton	7615	4001	1446
Luhn-DF	11132	5998	3577
TC	16540	10760	4689
TV	16540	10760	4689
TVQ	16540	10760	4689

Table 2. Attribute selection results

3.2. Attribute set validation

The evaluation of unsupervised tasks is a difficult problem due to the lack of objective measures. Here we have decided to work with labeled data sets in order to obtain a supervised efficiency measure about the generated datasets. For each data set generated in the attribute selection step, two classifiers (C4.5 decision tree and SVM) were constructed. Here we have used the WEKA environment [14] to induce these classifiers, adopting all software default parameters.

For each of these classifiers, their accuracy rate was estimated using 10-fold cross validation. It is important to emphasize that we are not interested in evaluating what classifier is better than the other; we just focused on comparing how the different feature

selection algorithms reacts under a same evaluation process. In Table 3, it is possible to see the evaluation results.

At the first glance, focusing only on the feature subsets performance, it is possible to see that Term Contribution, Term Variance and Term Variance Quality tend to present better results than Luhn TF, Salton and Luhn DF cutoffs. Although any statistical significance difference can be inferred, the constant better results presented by these methods (for example, TC has always presented better accuracy than TF and DF methods) and a subjective analysis of the eliminated attributes encouraged us to point that they are good choices to select representative attributes, even when dealing with composed terms.

		Data sets accuracy (%)		
Algorithms	**Cutoff**	**IFM**	**CBR-ILP-IR-SON**	**20 Newsgroups**
C4.5	**Luhn-TF**	82.13±3.84	84.26±4.72	38.57±2.79
	Luhn-DF	82.48±2.5	83.70±4.80	39.57±3.57
	Salton	76.10±5.04	80.56±5.34	38.58±3.85
	TC	83.33±3.44	88.48±4.04	39.77±3.71
	TV	83.33±3.72	88.50±4.04	37.98±3.24
	TVQ	83.33±3.72	88.48±4.04	40.48±2.72
SVM	**Luhn-TF**	78.68±4.06	94.11±2.80	37.18±3.59
	Luhn-DF	78.00±3.38	93.05±3.02	34.57±4.08
	Salton	74.04±4.56	91.95±3.25	34.37±2.64
	TC	78.85±3.79	95.00±2.70	38.48±3.67
	TV	78.51±4.14	95.06±2.59	37.98±3.72
	TVQ	78.85±4.29	95.11±2.50	40.28±3.94

Table 3. Attribute sets validation results

Another aspect that can be observed is the abrupt fall down of classifier accuracy from the IFM and CBR-ILP-IR-SON data sets to the Twenty Newsgroup data set. One possible reason is that the first two text collections are composed by scientific articles, so they were written using domain terms. Therefore, the newsgroup messages are not written neither reviewed under the same criteria, which implies in the use of poor or non specific vocabulary. So, the quality of the selected terms is higher in the first two text collections, improving the classifier efficiency.

4. Conclusions

Generating and selecting good subsets of attributes is not a trivial task. It demands careful and hard work, with no technical glamor. In this paper, some unsupervised methods to deal with these two problems were presented.

First of all, it is important to emphasize that the application of both statistical tests on the attribute generation and attribute selection methods are crucial to turn the knowledge extraction process computationally viable. Some experiments using all the generated terms on the three datasets used here were carried out and the high dimensionality of the attributes presented in this context and its consequently extremely high memory demand made it impossible to extract classification models.

Concerning the attribute generation, some methods that can be used to generate onegrams and twograms as terms were presented. These tests were applied attempting to select the most conceptually representative terms. In this sense, the log of the likelihood

ratio test was used. Analyzing the selected twograms, it was possible to see that this statistical cutoff has eliminated must of the non-sense ones, reducing in more than fifty per cent the number of the selected twograms. As this work is inserted in a process of topic taxonomy generation, the use of twogram terms is important because it can improve the comprehensibility of the final results.

Analyzing the attribute selection methods, it is possible to see that a good choice for unsupervised attribute selection is the use of variance-based selection algorithms. The methods presented here (Term Contribution, Term Variance and Term Variance Quality) showed a tendency to perform better than term and document frequency-based methods; besides, their easy implementation and low computational cost are attractive. Therefore, comparing the three variance-based methods, the difference between their efficiency is almost insignificant, being very similar in all three used data sets. Analyzing the terms selected by these three terms, it was possible to see that the three methods share almost 85% of selected terms in all three datasets, perhaps because of the nature of the text collections. This allows the text mining specialist to use the one that is more suitable to his work or even the computationally cheaper among them.

In future works, more techniques for both generation and selection of terms will be compared. Also, a subjective evaluation of the term set representativeness in each problem will be carried out by domain specialists. Finally, tests with a variation in the percentage of selected attributes of all the methods shown here will be applied, in order to deeply verify the difference among them.

References

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] S. Banerjee and T. Pedersen. The design, implementation, and use of the ngram statistics package. In A. F. Gelbukh, editor, *CICLing*, volume 2588 of *Lecture Notes in Computer Science*, pages 370–381. Springer, 2003.

[3] I. Dhillon, J. Kogan, and C. Nicholas. Feature selection and document clustering. In M. W. Berry, editor, *Survey of Text Mining*, pages 73–100. Springer, 2003.

[4] L. Liu, J. Kang, J. Yu, and Z. Wang. A comparative study on unsupervised feature selection methods for text clustering. *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*, pages 597–601, 30 Oct.-1 Nov. 2005.

[5] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma. An evaluation on feature selection for text clustering. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pages 488–495. AAAI Press, 2003.

[6] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal os Research and Development*, 2(2):159–165, 1958.

[7] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[8] E. T. Matsubara, C. A. Martins, and M. C. Monard. Pre-text: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Technical Report 209, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos, 2003.

[9] M. F. Moura. Uma abordagem para a construção e atualização de taxonomias de tópicos a partir de coleções de textos dinâmicas, 2006. Monografia de Qualificação de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos.

[10] M. Porter. An algorithm for suffixing stripping. *Program*, 14(3):130–137, July 1980.

[11] S. O. Rezende, J. B. Pugliesi, E. A. Melanda, and M. F. Paula. Mineração de dados. In S. O. Rezende, editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 12, pages 307–335. Manole, 1 edition, 2003.

[12] G. Salton and C. Yang. On the specification of term values in automatic indexing. 1973.

[13] G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Association Science*, 1(26):33–44, 1975.

[14] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, June 2005.