Intelligent Classification of Economic Activities from Free Text Descriptions

Elias Oliveira[1], **Patrick Marques Ciarelli**[2],
Wallace F. Henrique[3], **Lucas Veronese**[3] **Felipe Pedroni, Alberto F. De Souza**[3]

[1] Department of Information Science

[2]Department of Electric Engineering

[3]Department of Computer Science
Federal University of Espirito Santo
Av. Fernando Ferrari s/n
29060-970 – Vitoria, ES Brazil

{elias,alberto}@inf.ufes.br

Abstract. *We tackle the problem of automating the categorization of economic activities from business descriptions in free text format. This kind of information is vital to fundamental aspects of national governmental administration such as short, medium and long term planning and taxation. As the number of possible categories considered is very large (more than 1000 in the Brazilian scenario), the automatic text categorization problem targeted here is quite challenging. We have applied and compared the use of two different techniques to deal with it: the Vector Space Model in its classical form to represent the texts, and VG-RAM, a Weightless Neural Network.*

1. Introduction

Automatic text classification and clustering are still very challenging computational problems to the information retrieval (IR) communities both in academic and industrial contexts. Currently, the majority of the work on IR one can find in the literature is focused on classification and clustering of webpages. However, there are many other important applications to which little attention has hitherto been paid, which are as well very difficult to deal with. One example of these applications is the classification of companies based on their statements of purpose, also called mission statements, which represent the business context of the companies' activities.

The categorization of companies according to their economic activities constitute a very important step towards building tools for obtaining information for performing statistical analyses of the economic activities within a city or country. With this goal, the Brazilian government is creating a centralized digital library with the statement of purpose of all companies in the country. This library will serve the three government levels: Federal; the 27 States; and the more than 5.000 Brazilian counties. In order to categorize the statement of purpose of each Brazilian company, within this digital library, into the economic activities recognized by Brazilian law – more than 1.000 possible activities – we estimate that the data related to more than 5 million companies will have to be processed, and that at least 300.000 statements of purpose of new companies, or of companies which

are changing their statement of purpose, will have to be processed every year. It is important to note that the large number of possible categories makes this problem particularly complex when compared with others presented in the literature [Sebastiani 2002].

This work presents some preliminary experimental results on automatic categorization of a set of 3281 statements of purpose of Brazilian companies into a subset of 764 economic activities recognized by Brazilian law. We used two techniques in our experiments: Vector Space Model [Salton et al. 1975], for the representation of the statement of purposes documents, and Weightless Neural Networks (WNN) [Ludermir et al. 1999]. The former, the representation technique with the cosine as the similarity metric between any two documents, we will named it by (VS), was chosen because of its popularity in the IR literature and will serve also as a basis for the later technique. The best performing technique, weightless neural network, has shown 67.03% accuracy in identifying a correct category for each of the 3281 statements of purpose. To our knowledge, this is the first report on using WNN for text categorization into a large number of classes as that used in this work and the results are very encouraging.

This work is organized as follows. In Section 2, we point out some of the characteristics of the problem and its importance for the government institutions in Brazil. In Section 3 the preliminary experimental results are discussed. The two techniques, VS and WNN, were used and compared. We present our conclusions and indicate some future paths of this research.

2. The Problem

In many countries, companies must have a contract (*Articles of Incorporation* or *Corporate Charter*, in USA), with the society where they can legally operate. In Brazil, this contract is called a *social contract* and must contain the *statement of purpose* of the company – this statement of purpose must be categorized into a legal business activity by Brazilian government officials. For that, all legal business activities are cataloged using a table called CNAE – *Classificação Nacional de Atividade Econômicas* (National Classification of Economic Activities) [CNAE 2003].

To perform the categorization, the government officials (at the Federal, State and County levels) must find the semantic correspondence between the company statement of purpose and one or more entries of the CNAE table. There is a numerical code for each entry of the CNAE table and, in the categorization task, the government official attributes one or more of such codes (CNAE codes) to the company at hand. This can happen on the foundation of the company or in a change of its social contract, if that modifies its statement of purpose.

The computational problem addressed by us is that of finding automatically the semantic correspondence between a statement of purpose of a company and one or more items of the CNAE table. To do that, we have employed in this work two techniques: VS and WNN.

3. Experiments

Our dataset is organized so that we have two matrixes, C and D. Each element of a vector in C is the frequence of a relevant word present in the official brief textual description

(an average of 8 words and, in many cases, as small as two words) of each one of the 764 CNAE codes [CNAE 2003]. We removed from this set of words the Portuguese stopwords. The gender and plurals were also removed, but only the trivial cases of the Portuguese grammar.

The vectors $d_i \in D$ were built in a similar way, however, in this case we considered only the frequencies of those words which have already appeared in any of the $c_i \in C$ vector. Thus, each of the 3281 statements of purpose, an average of about 70 words each, was also represented as a vector with the occurrence frequencies of the words as its components. The result was a matrix C of 1001 lines and 764 columns, and a matrix D of 1001 by 3281 positions (elements).

The CNAE codes of each company in the dataset were assigned by trained Brazilian government officials. The number of codes assigned to each company varies from 1 to 12. We have considered an automatic assignment as correct when the technique under examination selected any of the classes assigned by the human specialist to that statement of purpose.

To categorize the 3281 statements of purpose into the 764 CNAE codes, both represented by the **vector space model**, the cosine similarity measure was used. Therefore we computed all the $cos(d_i, c_j), \forall c_j \in C$, and the largest cosine was selected as the category for d_i [Salton et al. 1975]. This strategy resembles the kNN algorithm, when $k = 1$ [Soucy and Mineau 2001].

Weightless Neural Network, also known as RAM-based neural networks, or n-tuple classifiers, employ neurons that operate with binary input values and use Random Access Memories (RAM) as lookup tables: the synapses of each neuron collect a vector of bits from the network's inputs that is used as the RAM address, and the value stored at this address is the neuron's output. Training can be made in one shot and basically consists of storing the desired output in the address associated with the input vector of the neuron [Aleksander 1996]. In this work we use Virtual Generalizing RAM (VG-RAM) networks, which are WNNs whose neurons store the input-output pairs shown during training [Aleksander 1998], instead of only the output. In the recall phase, the memory of VG-RAM neurons is searched associatively by comparing the input presented to the network with all inputs in the input-output pairs learned. The output of each VG-RAM neuron is taken from the pair whose input is nearest to the input presented – the distance function employed is the Hamming distance. If there is more than one pair at the same minimum distance from the input presented, the neuron's output is chosen randomly among these pairs.

To categorize texts using the VG-RAM WNN, we employed a network consisting of 121 (11x11) neurons with 512 synapses each. The synapses of the neurons, randomly connected to an input vector of 1001 elements, collect 1 or 0 from the input depending on whether it contains a value larger or equal to 0, respectively, while the output of the neurons can assume a value between 1 and 764. During training, the VG-RAM WNN input vector was fed with the columns of the first matrix, C, and the output with a value equal to the order of each column (an index to the CNAE table entry). During recall, the network was fed with each column of the second matrix, D, and all 121 outputs of the VG-RAM WNN were evaluated for each column. The value of the majority (the order of

the column of first matrix, learned during training) was taken as the network's output.

VS	VG-RAM WNN
63.36%	67.03%

Table 1. Percentage of correct CNAE code assignments of each technique.

Table 1 presents the categorization performance of each technique as a percentage of correct CNAE code assignments to the 3.264 statements of purpose. As Table 1 shows, VG-RAM WNN outperforms VS by 3.67%.

4. Conclusions

This paper presented an experimental evaluation of the performance of the VS and VG-RAM WNN techniques on automatic free text categorization into economic activities. We have trained the WNN system with 764 brief official Brazilian descriptions of economic activities and use them to categorize 3281 companies into these economic activities according to the statements of purpose of each one of these companies. Our experiments showed that WNN can outperform VS for a significant margin: 67.03% $\times$ 63.36% accuracy, respectively. It is important to note the large number of categories used in the experiments, 764. Besides, to the authors knowledge, this is the first report on using WNN for text categorization into a large number of classes.

As future work, one of the improvements we are working on for the VS algorithm is the use of the artificial centroid vector strategy for improving selectivity [Salton et al. 1975], while, for WNN, we are studying the use of knowledge correlation between the input-output pairs learned [Carneiro et al. 2006].

5. Acknowledgments

We would like to thank Andréa Pimenta Mesquita, CNAE classifications coordinator at Vitoria City Hall, for providing us with the dataset we used in this work. In addition, we would also like to thank Hannu Ahonen, Felipe M. G. França, Priscila Machado Vieira Lima and Eliana Zandonade for their technical support and valuable comments on this work. This work is partially supported by the Internal Revenue Brazilian Service (*Receita Federal do Brasil*) and the CNPq, the Brazilian government research agency, under the projects numbers: 134830/2006-7 of the second author; 131900/2006-4 of the fifth; and 504156/2004-7, 308207/2004-1, 471898/2004-0 and 620165/2006-5 of the last author.

References

Aleksander, I. (1996). Self-adaptive Universal Logic Circuits (Design Principles and Block Diagrams of Self-adaptive Universal Logic Circuit with Trainable Elements). *IEE Electronic Letters*, (2):231–232.

Aleksander, I. (1998). From WISARD to MAGNUS: a Family of Weightless Virtual Neural Machines. In *RAM-Based Neural Networks*, pages 18–30. J. Austin.

Carneiro, R., Dias, S. S., Fardin Jr., D., Oliveira, S., Garcez, A. S. d., and De Souza, A. F. (2006). Improving VG-RAM Neural Networks Performance Using Knowledge Correlation. *Lecture Notes on Computer Science*, 4232:427–436.

CNAE (2003). *Classificação Nacional de Atividades Econômicas Fiscal*. IBGE – Intituto Brasileiro de Geografia e Estatística, Rio de Janeiro, RJ, 1.1 edition. http://www.ibge.gov.br/concla.

Ludermir, T. B., Carvalho, A. C. P. L. F., Braga, A. P., and Souto, M. d. (1999). Weightless Neural Models: a Review of Current and Past Works. *Neural Computing Surveys*, 2:41–61.

Salton, G., Wong, A., and Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11):613–620.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Soucy, P. and Mineau, G. W. (2001). A Simple KNN Algorithm for Text Categorization. In *ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 647–648, Washington, DC, USA. IEEE Computer Society.