Automatic Text Categorization: Case Study

Renato Fernandes Corrêa, Teresa Bernarda Ludermir
Centro Informática da UFPE
E-mails: rfc@cin.ufpe.br, tbl@cin.ufpe.br

Text Categorization is a process of classifying documents with regard to a group of one or more existent categories [1] according to themes or concepts present in their contents. The most common application of it is in Information Retrieval Systems (IRS) to document indexing [2].

The organization of text in categories allow the user to limit the target of a search submitted to IRS, to explore the collection and to find relevant information to they need with poor knowledge about the keywords of a theme.

A method to transform text categorization into a viable task is to use machine-learning algorithms to automate text classification, allowing it to be carried out fast, into concise manner and in broad range.

The objective of this work is to present and compare the results of experiments on text categorization using artificial neural networks of the type Multilayer Perceptron (MLP) [3] and Self-organizing Maps (SOM) [6], and traditional machine-learning algorithms used in this task [4]: C4.5 decision tree, PART decision rules and Naive Bayes classifier.

The experiments were carried out with three collections of texts, the collection K1 [5], the collection PubsFinder [4] and a subcollection of the Reuters-21758 Collection called Metals Collection [1].

Comparing the best performance of each algorithm, in terms of classification error on test set for each collection, the experimental results show artificial neural networks as good classifiers for problems of text categorization. In general, the MLP Networks distinguished as the bests classifiers and the SOM networks had better performance than the symbolic machine learning algorithms.

The classification error obtained by SOM was not twice bigger than the minor founded by the other classifiers for the collections. Thus, SOM networks can be used as an auxiliary tool to manual text categorization, as well as a way to explore a text collection, having as initial interface the map generated and labeled with the most numerous category in each neuron.

References:

[1] C. B. Rizzi, J. F. Valiati and P. M. Engel, "Uma Proposta para Categorização de Textos por uma Rede Neural", *Proceedings of V Congresso Brasileiro de Redes Neurais- VI Escola de Redes Neurais*, Rio de Janeiro, 2001, pp. 517-522.

[2] R. F. Corrêa. *Categorização de Documentos utilizando Redes Neurais: Análise comparativa com técnicas não-conexionistas*. Dissertação de Mestrado. Centro de Informática da UFPE, Recife, 2002.

[3] A. P. Braga, T. B. Ludermir and A. C. P. L. F. Carvalho, *Redes Neurais Artificiais: Teoria e aplicações*, LTC - Livros Técnicos e Científicos Editora S.A, Rio de Janeiro, 2000.

[4] M. L. Neves, *PubsFinder - um Agente Inteligente para Busca e Classificação de Páginas de Publicações*. Dissertação de Mestrado. Centro de Informática da UFPE, Recife, 2001.

[5] D. Boley, M. Gini, R. Gross, E. Han, K. Hastings, G. Karypis, V. Kumar, B. Mobasher and J. Moore, "Partitioning-based clustering for web document categorization", *Decision Support Systems,* v.27, 1999, pp. 329-341.

[6] T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela and V. Paatero and A. Saarela, "Self Organization of a Massive Document Collection", *IEEE Transaction on Neural Networks*, v. 11, n. 3, May 2000.

Proceedings of the VII Brazilian Symposium on Neural Networks (SBRN'02)
0-7695-1709-9/02 $17.00 © 2002 **IEEE**