Filtering Very Similar Text Documents: A Case Study

Jiří Hroza[1], Jan Žižka[1], and Aleš Bourek[2]

[1] Faculty of Informatics, Department of Information Technologies
Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
{xhroza1,zizka}@informatics.muni.cz
[2] Faculty of Medicine, Department of Biophysics
Masaryk University, Joštova 10, 662 43 Brno, Czech Republic
bourek@med.muni.cz

Abstract. This paper describes problems with classification and filtration of similar relevant and irrelevant real medical documents from one very specific domain, obtained from the Internet resources. Besides the similarity, the documents are often unbalanced—a lack of irrelevant documents for the training. A definition of similarity is suggested. For the classification, six algorithms are tested from the document similarity point of view. The best results are provided by the back propagation-based neural network and by the radial basis function-based support vector machine.

1 Introduction

After downloading many textual documents from the Internet resources, users often need subsequent filtration of the resulting data. Typically, only a small part of these documents is relevant for a user. If the documents are sufficiently different, it is possible to use an effective filtering method like, for example, the naïve Bayes classifier provided that there are balanced and good training sets of instances, see [6]. However, if a user looks for documents from a very specific and narrow area and moreover he or she defines efficient search conditions, the outcome of a web-browser searching can include very similar relevant and irrelevant documents. The similarity is usually based on a high incidence of identical words so a classifier can make many errors when filtering irrelevant documents even if it is trained using carefully selected examples. An additional problem is often in very unbalanced numbers of training positive and negative examples because users usually need processing of data that they simply obtained and it is not possible to create more positive or negative examples to fulfil conditions required by algorithms for the reliable training, see [9]. Thus the existing methods based on extensive balanced training, such as [4], or on general-purpose onthologies, such as [5], are not efficient for similar special-purpose texts.

The described situation occurs very often in various medical domains and this paper describes a study with documents in the area of gynecology, infertility, and assisted reproduction. Any suitable solution of the problem is naturally