Selecting Interesting Articles Using Their Similarity Based Only on Positive Examples

Jiří Hroza and Jan Žižka

Faculty of Informatics, Department of Information Technologies
Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
{xhroza1,zizka}@informatics.muni.cz

Abstract. The task of automated searching for interesting text documents frequently suffers from a very poor balance among documents representing both positive and negative examples or from one completely missing class. This paper suggests the *ranking approach* based on the k-NN algorithm adapted for determining the *similarity degree* of new documents just to the representative *positive* collection. From the viewpoint of the precision-recall relation, a user can decide in advance how many and how similar articles should be released through a filter.

1 Introduction

When selecting from unstructured natural language text documents, a pragmatic trouble can aggravate the design of a filter: many users collect articles that represent (almost) only the interesting ones, and the required *relevant negative examples* for training an algorithm are missing. Typically physicians, having only positive examples of articles, need to automatically single out very specific medical documents within a narrow expert area—yet, containing too many articles around very similar topics [1]; here is the inspiration for the described research. The problem with synthetical filling in the missing examples is that *arbitrary* text documents different from the positive ones cannot be generally used: how to define effectively the dissimilarity? This paper describes the *ranking approach* based on the k-NN (k-nearest neighbors) algorithm adapted for determining the similarity of articles to the representative *positive* examples. For the comparison, outcomes of the SVM (*support vector machines*) algorithm are also shown.

2 Text Documents and Their Preprocessing

To test performance of the one-class k-NN and SVM, one of the standard benchmarks 20Newsgroups dataset was used[1]. Then, the one-class k-NN was also applied to a specific set of real expert medical documents[2] from MEDLINE [1]. Porter's algorithm [4] was applied to obtain a stem of each word. The dictionary was created as a set of all distinct words in the exemplary articles (*bag of*

[1] http://www.ai.mit.edu/~jrennie/20Newsgroups/

[2] http://www.fi.muni.cz/~xhroza1/datasets/glall/