Chapter VIII

Text Processing by Binary Neural Networks

T. Beran, Czech Technical University, Czech Republic
T. Macek, Czech Technical University, Czech Republic

ABSTRACT

This chapter describes a rather less traditional technique of text processing. The technique is based on the binary neural network Correlation Matrix Memory. We propose using the neural network for text searching tasks. Two methods of coding input words are described and tested. Further, we discuss the problems of using this approach for text processing.

INTRODUCTION

With more and more people becoming familiar with computers, the amount of information stored in electronic formats is quickly increasing. The consequence of that is the need to be able to search large amounts of data for particular information. Various techniques have been developed for the text searching task. Many techniques are very fast and sophisticated. Speed is one of the most important criteria, but it is not the only one. The other one is the
ability to deal with somehow corrupted text. Text could be corrupted, for example, when we do not know exactly what we are searching for, or if the text is the result of OCR or speech recognition.

In this chapter, we describe a rather less traditional technique for the text searching task. The technique is based on a binary neural network called CMM (Correlation Matrix Memory). We have tested CMM on the problem of finding a particular word in a single text document. Searching gives all occurrences of the word. Although the technique is able to search approximately, here we focus on exact searching.

This chapter is divided as follows: “Technique Description” explains our approach and describes CMM; “Text Coding and Experiments” explains the importance of input patterns coding, proposes two new methods, and shows initial experiments; “Discussion” discusses some problems that arise when this technique is applied to real text.

TECHNIQUE DESCRIPTION

In this section, we describe the principle of the technique based on binary neural network as well as on CMM itself. Usually, a fast conventional technique is used for text searching tasks, e.g., Boyer-Moor or Shift-or algorithms. Conventional techniques go through text and compare it with a searched pattern. Our technique is based on storing associations between words and their location in the text. The association is stored in CMM. This approach is similar to techniques such as inverted file lists or hash tables (Hodge & Austin, 2001).

Our approach is shown in Figure 1. It operates in two phases — the learning and the recalling. The learning phase stores associations between words and their positions in the text document. The recalling phase searches for a query word. In fact, there is no search during the recalling phase; it simply recalls a proper association.

In the learning phase, the parser goes through the text and cuts off the words. All words are coded and pass their code word to the input of the CMM. Coding input words significantly influences properties of CMM and the recall process. It checks whether the input word has been already learned. If it has, the word is learned no more; but its position in the text is added to a proper entry of the position table. Otherwise, a code for this word is produced and put at the output of CMM. Then, a new association is learned, and a new entry in the position table is added. The position table contains entries for all learned words.
An Efficient Approach for Incremental Association Rule Mining through Histogram Matching Technique
[www.igi-global.com/article/efficient-approach-incremental-association-rule/74782?camid=4v1a](www.igi-global.com/article/efficient-approach-incremental-association-rule/74782?camid=4v1a)

Application of Fuzzy User's Profile for Mining Reusable E-Learning Repositories on Web through Lightweight Mobile Agent
[www.igi-global.com/chapter/application-of-fuzzy-users-profile-for-mining-reusable-e-learning-repositories-on-web-through-lightweight-mobile-agent/138713?camid=4v1a](www.igi-global.com/chapter/application-of-fuzzy-users-profile-for-mining-reusable-e-learning-repositories-on-web-through-lightweight-mobile-agent/138713?camid=4v1a)