Approximate text searching

Gonzalo Navarro

Departamento de Ciencias de la Computación
Universidad de Chile

Advisor: Ricardo Baeza-Yates

December 1998

Abstract

This thesis focuses on the problem of text retrieval allowing errors, also called "approximate" string matching. The problem is to find a pattern in a text, where the pattern and the text may have "errors". This problem has received a lot of attention in recent years because of its applications in many areas, such as information retrieval, computational biology and signal processing, to name a few.

The aim of this work is the development and analysis of novel algorithms to deal with the problem under various conditions, as well as a better understanding of the problem itself and its statistical behavior. Although our results are valid in many different areas, we focus our attention on typical text searching for information retrieval applications. This makes some ranges of values for the parameters of the problem more interesting than others.

We have divided this presentation in two parts. The first one deals with on-line approximate string matching, i.e. when there is no time or space to preprocess the text. These algorithms are the core of off-line algorithms as well. On-line searching is the area of the problem where better algorithms existed. We have obtained new bounds for the probability of an approximate match of a pattern in a random text, and used these results to analyze many old and new algorithms. We have developed new algorithms for this problem which are currently among the fastest known ones, being even the fastest algorithms for almost all the interesting cases of typical text searching. Finally, we extended our results to the simultaneous search of multiple patterns, obtaining the best existing algorithms when a moderate number of them is sought (less than 100,

approximately).

The second part of this thesis addresses indexed approximate string matching, i.e. when we are able to build an index for the text beforehand, to speed up the search later. The ultimate index for approximate string matching is yet to appear and the current development is rather immature, but we have made progress regarding new algorithms as well as better understanding of the problem. For the restricted case of indices able to retrieve only whole words on natural language text, we have obtained new analytical results on their asymptotic complexity, which allowed us to develop an index that is sublinear in space and query time simultaneously, something that did not exist before. For this kind of index we also presented improved search algorithms. For general indices able to find any occurrence (not only words), we have developed new indexing schemes which are a tradeoff between efficiency and space requirements. Also, inspired in on-line techniques, we have proposed a hybrid between existing indexing schemes and obtained very promising results.

It is worth to mention that in almost all cases we have complemented the development of the new algorithms with their worst-case and average-case complexity analysis, as well as a thorough experimental validation and comparison against the best previous work we were aware of.

As a whole, we believe that this work constitutes a valuable contribution to the development and understanding of the problem of approximate text searching.