

New Indices for Text: PAT trees and PAT arrays

Gaston H. Gonnet
Dept. of Computer Science
ETH, Zurich
Switzerland

Ricardo A. Baeza-Yates
Depto. de Ciencias de la Computación
Universidad de Chile
Casilla 2777,
Santiago, Chile

Tim Snider
Centre for the New OED and Text Research
University of Waterloo
Waterloo, Ontario
Canada N2L 3G1

Abstract

We survey new indices for text, with emphasis on PAT arrays (also called suffix arrays). A PAT array is an index based on a new model of text which does not use the concept of word and does not need to know the structure of the text.

1 Introduction

Text searching methods may be classified as lexicographical indices (indices that are sorted), clustering techniques, and indices based on hashing. In this chapter we discuss two new lexicographical indices for text, called PAT trees and PAT arrays. Our aim is to build an index for the text of size similar to or smaller than the text.

Briefly, the traditional model of text used in information retrieval is that of a *set of documents*. Each document is assigned a list of *keywords* (attributes), with optional relevance *weights* associated