Invited Talk

Current Challenges in Textual Databases

Gonzalo Navarro, Universidad de Chile, Chile

The basic text search problem consists in finding the occurrences of a short string, called the pattern, inside a long string, called the text. More complex variants permit that the pattern contains optional or repeatable characters, gaps, and other sophistications until regular expression searching. It is also possible that one wants to find occurrences that are similar enough to the pattern.

This is one of the best studied problems in Computer Science, since it appears in diverse forms in almost every area. In particular, it is a central problem in areas such as information retrieval, signal processing, computational biology, and textual and multimedia databases. Solutions to the problem, however, are rather different depending on the nature of the application and the type of the pattern to search for.

In this talk we will focus on the problem of managing large textual databases where the text is general. This means that the text cannot be treated as natural language (i.e., formed by words) but as a sequence of characters: every substring of the text can be retrieved. This abstracts from different text retrieval problems in Oriental or agglutinating languages, searching in musical, genetic and proteic sequences, and it can even be relevant in Western text searching. On the other hand, the size of those texts makes sequential searching unaffordable: it is necessary to build an index so as to speed up searches. Finally, it is necessary to permit updates to the text and to maintain the index up to date.

Although there exist solutions to the general text indexing problem from a long time ago, these solutions can never be actually applied to serious cases of real life, because of three reasons: (1) indexes are huge (4 to 30 times the text size); (2) indexes cannot be modified to reflect changes in the text, but they must be reconstructed from scratch; (3) indexes do not behave well in secondary memory, which, given their large size, makes them useless except for very small texts. The current state of the art can be described as rather primitive. However, in the last years much effort has been done in this area.

In this talk we will review recent efforts to solve the three aforementioned problems, and to obtain a real solution to large general textual databases. On one hand, research is being done on compressed text indexes, which contain and replace the text, permit fast searching and displaying of any portion of the text, and yet they take less space than the original text. On the other hand, there exist several data structure proposals that can be efficiently updated when the text changes. Finally, we will review secondary memory schemes to build and store indexes on disk, trying to minimize the number of disk pages accessed and making good use of the space in disk blocks.

Short Biography

Gonzalo Navarro obtained his PhD in Computer Science at the University of Chile in 1998, and in early 1999 became Assistant Professor. Between 1999 and 2000 he made a postdoctoral stay at University of Helsinki (Finland) and Institut Gaspard Monge, University of Marne-la-Valle (France). In 2003 he became Associate Professor at University of Chile.

Since 1994 he has researched in several areas related to design and analysis of algorithms, textual databases and approximate searching. His current research interests are: exact and approximate, sequential and indexed search for simple and complex patterns in linear or structured, plain or compressed text, including in particular compressed text indexes; and metric space searching. He has coauthored a book, five book chapters, more than 30 articles in international journals, and more than 60 articles in international conferences.