Abstract

Strings play a fundamental role in computer science. Data is codified into strings, and by interpreting them information can be derived. Given a set of strings, few interesting questions arise, such as *"are these strings related?"*, and *"if they are related, how can we measure this relatedness?"*. The definition of a degree of similarity (or correlation) between strings is strongly important. Different definitions of similarity between strings already exist in literature and they steam from the concept of *metric* in mathematics. One of the most famous and well-known string similarity metric is the *edit distance*, which measures the minimum number of *edit operations* required to transform one string into another one. However, in the definition of the similarity between two strings, one important natural assumption is made: identical symbols among strings represent identical information, whereas different symbols introduce some form of differentiation. This last assumption results to be extremely reductive. In fact, there are cases in which symbol identity seems to be not enough, and even if there are no common symbols between two strings, it could happen that they represent similar information. Moreover, there are cases in which a one-to-one mapping between symbols is not enough, thus a *many-to-many* mapping is needed. The necessity of a suitable metric capable of capturing *hidden* correlations between strings emerges and this metric should take into account that *different* symbols may express *similar* concepts.

This thesis aims to provide a contribution in this setting. Initially, we present a framework that generalizes most of the existing string metrics based on symbol identity, making them suitable for application scenarios which involve strings defined on heterogeneous alphabets. We formally define the Multi-Parameterized Edit Distance, a generalization of the edit distance with the support of our framework, and we discuss its computational issues.

Then, we present various heuristics designed, implemented and tested out, in order to approach computational issues of the generalization: we start with a survey on heuristics to acquire a global view of the problem, then we select, discuss and test three of them in detail.

In the last part, we discuss several application contexts which have been studied in this thesis. These scenarios span from engineering to biomedical informatics. In particular, they concentrate on Wireless Sensors Area Networks, White Matter Fiber-Bundles analysis and Electroencephalogram analysis.

Finally, at the end of the thesis, we draw our conclusions and highlight future work.