Practical Text Mining

Ronen Feldman

Data Mining Laboratory
Department of Mathematics and Computer Science
Bar-Ilan University
Ramat-Gan Israel 52900
feldman@cs.biu.ac.il

Abstract

Knowledge Discovery in Databases (KDD) focuses on the computerized exploration of large amounts of data and on the discovery of interesting patterns within them. While most work on KDD has been concerned with structured databases, there has been little work on handling the huge amount of information that is available only in unstructured textual form. In this tutorial we will present the general theory of Text Mining and will demonstrate several systems that use these principles to enable interactive exploration of large textual collections. We view Text Mining as a combination of Information Retrieval methods and Data Mining methods. We will describe generic techniques for text categorization and information extraction that are used by these systems. The systems that will be presented are KDT which is system for Knowledge Discovery in Texts, FACT, which discovers associations amongst keywords labeling the items in a collection of textual documents, and Text Explorer which is a system that provides a high level language for interactive exploration of textual collections. We will present a general architecture for text mining and will outline the algorithms and data structures behind the systems. We will give special emphasis to incremental algorithms and to efficient data structures. The Tutorial will cover the state of the art in this rapidly growing area of research.