IDEA GROUP PUBLISHING **ITB13683**

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

This paper appears in the publication,
Artificial Intelligence and Integrated Intelligent Information Systems: Emerging Technologies and Applications
edited by X. Zha © 2007, Idea Group Inc.

Chapter VI

Intelligent Text Mining: Putting Evolutionary Methods and Language Technologies Together

John Atkinson, Universidad de Concepción, Chile

Abstract

This chapter introduces a novel evolutionary model for intelligent text mining. The model deals with issues concerning shallow text representation and processing for mining purposes in an integrated way. Its aims are to look for interesting explanatory knowledge across text documents. The approach uses natural-language technology and genetic algorithms to produce explanatory novel hidden patterns. The proposed approach involves a mixture of different techniques from evolutionary computation and other kinds of text mining methods. Accordingly, new kinds of genetic operations suitable for text mining are proposed. Some experiments and results and their assessment by human experts are discussed which indicate the plausibility of the model for effective knowledge discovery from texts. With this chapter, authors hope the readers to understand the principles, theoretical foundations, implications, and challenges of a promising linguistically motivated approach to text mining.

Introduction

Like gold, information is both an object of desire and a medium of exchange. Also like gold, it is rarely found just lying about. It must be mined, and as it stands, a large portion of the world's electronic information exists as numerical data. Data mining technology can be used for the purpose of extracting "nuggets" from well-structured collections that exist in relational databases and data warehouses. However, 80% of this portion exists as text and is rarely looked at: letters from customers, e-mail correspondence, technical documentation, contracts, patents, and so forth.

An important problem is that information in this unstructured form is not readily accessible to be used by computers. This has been written for human readers and requires, when feasible, some natural language interpretation. Although full processing is still out of reach with current technology, there are tools using basic pattern recognition techniques and heuristics that are capable of extracting valuable information from free text based on the elements contained in it (e.g., keywords). This technology is usually referred to as text mining and aims at discovering unseen and interesting patterns in textual databases.

These discoveries are useless unless they contribute valuable knowledge for users who make strategic decisions (i.e., managers, scientists, businessmen). This leads then to a complicated activity referred to as knowledge discovery from texts (KDT) which, like knowledge discovery from databases (KDD), correspond to "the non-trivial process of identifying valid, novel, useful, and understandable patterns in data."

Despite the large amount of research over the last few years, only few research efforts worldwide have realised the need for high-level representations (i.e., not just keywords), for taking advantage of linguistic knowledge, and for specific purpose ways of producing and assessing the unseen knowledge. The rest of the effort has concentrated on doing text mining from an information retrieval (IR) perspective and so both representation (keyword based) and data analysis are restricted.

The most sophisticated approaches to text mining or KDT are characterised by an intensive use of external electronic resources including ontologies, thesauri, and so forth, which highly restricts the application of the unseen patterns to be discovered and their domain independence. In addition, the systems so produced have few metrics (or none at all) which allow them to establish whether the patterns are interesting and novel.

In terms of data mining techniques, genetic algorithms (GA) for mining purposes has several promising advantages over the usual learning / analysis methods employed in KDT: the ability to perform global search (traditional approaches deal with predefined patterns and restricted scope), the exploration of solutions in parallel, the robustness to cope with noisy and missing data (something critical in dealing with text information as partial text analysis techniques may lead to imprecise outcome data), and the ability to assess the goodness of the solutions as they are produced.

In order to deal with these issues, many current KDT approaches show a tendency to start using more structured or deeper representations than just keywords to perform further analysis so to discover informative and (hopefully) unseen patterns. Some of these approaches attempt to provide specific contexts for discovered patterns (e.g., "it is very likely that if X and Y occur then Z happens."), whereas others use external resources (lexicons, ontolo-