Preface

Text mining (TM) has evolved from the simple word processing at the end of the 1990s to now, when the concepts processing or even the knowledge extraction from linguistic structures are possible due to the most recent advances in its realm. The complexity inherent to the enormous and increasing amount of textual data can now be effectively approached, what enables the discovery of interesting relations in documents, e-mails, Web pages, and other nonstructured information sources. The need for effective TM approaches becomes more important if we consider that, according to the experts, most of the organizational information available in the modern enterprises comes in textual format.

TM can be defined as the application of computational methods and techniques over textual data in order to find relevant and intrinsic information and previously unknown knowledge. TM techniques can be organized into four categories: classification, association analysis, information extraction, and clustering. Classification techniques consist of the allocation of objects into predefined classes or categories. Association analysis can be applied to help the identification of words or concepts that occur together and to understand the content of a document or a set of documents. Information extraction techniques are able to find relevant data or expressions inside documents. Clustering is applied to discover underlying structures in a set of documents.

This book aims to provide researchers, graduate students, and practitioners the most recent advances in the TM process. The topics approached include: methodological issues, techniques, and models, successful experiences in applying this technology to real world problems, and new trends in this field. By offering this book to researcher and developer audiences, the editors aim to provide some start points for new developments in TM, and its correlated fields and bring to practitioners the newest information for building and deploying applications in the organizational context. Due to the meticulous process for selecting chapters,based on a strict peer-to-peer blind review, scholars can find interesting research results in the state of the art of Text Mining.

Critical applications can be developed in the fields of organizational and competitive intelligence, environmental scanning, prospective scenarios analysis, and business intelligence, among others, allowing the organizations to have a much more competitive position in the modern economy, while also obtaining a self knowledge with respect to their structure and dynamics. By applying the methods described in this book, one can transform textual patterns into valuable knowledge. This book goes beyond simply showing techniques to generate patterns from texts; it gives the road map to guide the subjective task of patterns interpretation.

Chapter I discusses information extraction related problems and provides an overview on the existing methodologies facing these problems. Rule learning based, classification based, and sequential labeling based methods are explained in detail. This discussion, along with some extraction applications experienced by the authors and examples of well-known applications in this field makes this chapter an excellent reference for a tutorial program.

Departing from a sound point of view about the importance of the information treatment and analysis as the most critical organizational competence for the information society, **Chapter II** brings an interesting discussion regarding the use of strategic information in competitive intelligence. Authors present a comprehensive review on the methodologies tuned to the French School of Information Science and discuss the necessary integration of information for analysis and the data and text mining processes, presenting examples of real-world applications.

Chapter III takes into account the competitive intelligence needs of gathering information related to external environment, analyzing, and disseminating the results to the board of a company. The authors hold with a disclosure tendency that leads companies to be more transparent as a demand posed by apprehensive shareholders in the face of unexpected crises in big companies. The chapter discusses natural language processing (NLP) techniques applied to analyze the increasing amount of organizational information published in that scenario, looking for competitive advantages.

NLP technology for mining textual resources and extracting knowledge are also the focus of **Chapter IV**. An overview on the NLP techniques involved, beyond a brief explanation on the application of these techniques, is provided along with two case studies. The case studies focus on finding definitions in vast text collections and obtaining a short multidocument summary of a cluster to respond to a relevant question posed by an analyst.

Chapter V describes an approach for deriving a taxonomy from a set of documents by using semantic information generated at the document sentence level. This information is obtained from a document profile model built by applying a frequent word sequence method. The agglomerative hierarchical method is applied to create a clustering configuration from which the taxonomy is derived. Experimental results show that the derived taxonomy can demonstrate how the information is interwoven in a comprehensive and concise form.

Chapter VI brings an original approach to knowledge discovery from textual data based on a fuzzy decision tree. The methods involved in the rule extraction process are described, including the inductive learning method, and the inference method involved. Two applications, one for analyzing daily business reports and other for rule discovering from e-mails, are presented in detail in order to make evident the effectiveness of the proposal.

Chapter VII extends the semisupervised multiview learning techniques for TM. Semisupervised learning is usually applied to build classification models over labeled and unlabeled data. The chapter contains the description of some algorithms related to this task, a review of applications found in literature, including Web page classification and e-mail classification, and experimental results found by analyzing academic and newsgroup related texts.

Chapter VIII proposes a multi-neural-network-agent-based Web text mining system for decision support by processing Web pages. The system comprises the processes of Web document searching, Web text processing, text feature conversion, and the neural model building. The scalability of this system is assured by the learning process implemented in a multi-agent structure. For performance evaluation, the authors describe a practical application for the crude oil price movement direction prediction.

Chapter IX focuses on a problem usually experienced by millions of Web users when searching for information: how to analyze hundreds or thousands of ranked documents. Research results have shown that if nothing interesting appears in the two or three top ranked documents, users prefer to reformulate their query instead of sifting through the numerous pages of search results that can hide relevant information. In this chapter an innovative approach based on the suffix tree clustering algorithm to sift the search results is proposed. Also, a review on other text mining techniques used in exploratory search solutions, a discussion on some advantages and disadvantages of contextualized clustering, and experimental results are presented.

Still in the Web search domain, **Chapter X** presents a study on the application of the ant colony optimization model to improve Web navigation. Considering the existing similarities in behavior between Web users' navigation and ant colonies, the chapter describes a system able to shorten the paths to a target page by applying information related to previous navigation. Although unable to have global vision due to their physical characteristics, ants at foraging use a curious marking system, in which any ant can leave a pheromone mark for others. The specific trails with stronger concentration of pheromone can guide the ants in finding shorter paths between their nest and the food source. AntWeb, the system resulting from the application of this metaphor, is described along with experimental results that demonstrate its benefits for users navigating in the Web and also its capacity to modify the Web structure adaptively.

Aiming to overcome the semantic mistakes that arise when applying the traditional word-based documents clustering techniques, **Chapter XI** introduces an alternative to cluster textual documents using concepts. The use of concepts instead of simple words adds semantics to the document clustering process, leading to a better understanding of large document collections, easing the summarization of each cluster and the identification of its contents. A useful methodology for document clustering and some practical experiments in a case study are shown in detail to demonstrate the effectiveness of the proposed alternative.

Chapter XII studies the text categorization (TC) techniques applied to the patent categorization problem. The patent categorization problem presents specific characteristics like the existence of many vague or general terms, acronyms, and new terms that make the traditional TC methods insufficient for an adequate categorization. Taking into account the application domain particularities and an important drawback in the existing TC solutions, the authors propose their method. An evaluation of this method on two English patent databases is presented and compared with other text categorization methods, showing how the proposed method outperforms the latter.

Chapter XIII presents an application of TM in the healthy management field. Nominal data related to patient medical conditions are used to define a patient severity index to model the quality of healthcare providers. A clustering analysis to rank groups of levels in the nominal data is carried out, issuing results that can help improving the healthcare system.

Finally, **Chapter XIV** enlightens an important problem of both data and text mining that usually remains unsolved due to their strong subjective nature: the clustering interpretation process. In order to deal with this inevitable characteristic of clustering analysis, that is, the hard burden of subjectivity, the authors built from the ontology of language a method to help the specialists in searching for a consensus when analyzing a given clustering configuration. This method, grounded in speech acts like affirmations, declarations, assessments, and so on, fits naturally as a pathway for the interpretation phase of Text Mining.

Hércules Antonio do Prado, DSc, Catholic University of Brasilia & Embrapa Food Technology
Edilson Ferneda, DSc, Catholic University of Brasilia
Editors
June 2007