Preface

In the information era, while the amount of electronic documents keeps growing exponentially, the time available to the final users to process the information tends to decrease. Moreover, text documents represent a powerful resource to infer new knowledge by means of data mining techniques. For instance, by analyzing the document content, it is possible to extract user interests, community opinions, and expand the knowledge about a specific domain. To help users quickly gain knowledge and ease the discovery of new information from text documents, a significant research effort has been devoted to the study and the development of automated summarization tools, which produce a concise overview of the most relevant document content (i.e., a summary). Extracting a succinct and informative description of document collections is fundamental to allowing users to quickly familiarize themselves with the information of interest. For example, the summary of a collection of news documents regarding the same topic may provide a synthetic overview of the most relevant news facets. Differently, the summarization of social network data can support the identification of relevant information about a specific event, and the inference of user and community interests and opinions.

During recent years, a considerable number of automated summarization tools have been proposed. For instance, summarizers have been developed oriented to producing concise representations of a single document and/or a collection of documents discussing the same topics. Both general-purpose and domain-oriented approaches may be identified. In addition, summarization systems may be classified as: (1) sentence-based, if they partition documents into sentences and select the most informative ones for inclusion in the summary, or (2) keyword-based, if they detect salient keywords that summarize the document content. All these approaches usually exploit statistical metrics, text mining algorithms and/or Natural Language Processing (NLP) methods. For instance, clustering-based approaches, probabilistic or co-occurrence-based strategies, graph-based algorithms, and itemset-based methods have already been proposed. Some systems also allow the management and the analysis of text documents in different languages. However, the evaluation of summaries represents a significant hurdle and can be very expensive in terms of time and resources. Thus, it is always difficult to identify which system can achieve the best results in the context of interest. The definition of automatic evaluation systems for text summaries is still an open research task.

This book provides a comprehensive discussion of the state of the art in document summarization. The reader will find in-depth discussion of the approaches focused on multilingual, domain-oriented and Web-oriented summarization tasks. Furthermore, some current real-world applications in several fields, such as contextual advertising, social networks, and archeology, are also provided.

The audience for the book includes—but is not limited to—students, lecturers, researchers, and practitioners of information technology, computer science, and bioinformatics. Developers and consumers

will be interested in discovering how document summarization can improve their productivity in real applications, while researchers and students can learn more about the main concepts of the state-of-the-art document summarization approaches. This book can also make an academic reference work by providing comprehensive coverage of the document summarization field. Through predictions of future trends, analysis of techniques and technologies, and focuses on real application scenarios, this book will be a useful instrument for developing new document summarization tools suited to different application fields.

The book comprises 13 chapters and is organized as follows.

Section 1, "General-Purpose and Domain-Specific Methods," consists of 5 chapters and provides a good overview of the data representation and evaluation metrics adopted by general-purpose and domain-specific summarizers.

Chapter 1, "Classification of Sentence Ranking Methods for Multi-Document Summarization," overviews state-of-the-art and most successful works focused on sentence-ranking methods. Since multi-document summarization approaches extract a subset of sentences that contain the most relevant information, the sentence-ranking task is fundamental to defining their importance for inclusion in a summary according to a relevance score. A categorization of ranking methods is also provided to highlight the different properties of each group.

Chapter 2, "Multi-Document Summarization by Extended Graph Text Representation and Importance Refinement," proposes the SentRel (Sentence Relations) method based on recursive importance inference from a three-tiered graph representation of a document collection and its refinement in the process of summary construction. Differently from other methods, the graph representation is not used to encode the relationships among terms, concepts, and/or sentences, but to capture different correlations among the documents in the collection. The approach is both language-independent and domain-independent and achieves good results with respect to state-of-the-art summarization systems on the TAC 2011 dataset.

Chapter 3, "Efficient Summarization with Polytopes," defines a system of linear inequalities to describe the content of the given document set. Each sentence is modeled by a hyperplane, and the intersections between these hyperplanes represent all possible summaries. Since a summary should preserve the content and meaning of the original document collection, the summary extraction task is re-formulated as finding the point on a convex polytope closest to the given hyperplane, which can be solved efficiently with linear programming. Experimental results on the DUC 2002 and MultiLing 2013 collections show good performance with respect to the other competitors.

Chapter 4, "Interactive Summaries by Multi-Pole Information Extraction for the Archaeological Domain," presents the application of a document summarization approach in a specific domain, the archeological one. The chapter proposes the Interactive Summary Extractor Tool (ISET), whose aim is to extract and organize text summaries for archeological and historical documental sources. This system is integrated in the Herodotus tool that supports domain experts in the reasoning tasks over complex interactions characterizing a society, in order to explain causes of events and predict future events according to some factor changes. The goal of the ISET tool is to improve summarization results according to user interests, thus simplifying the interaction with the system.

Chapter 5, "Evaluation Metrics for the Summarization Task," provides a comprehensive survey about several summary evaluation metrics. Indeed, the evaluation of summary quality is a challenging task, since, in many cases, it is subjective, costly, time consuming, and, if human-assisted, it can generate some bias. For these reasons, several works have targeted the definition of metrics and instruments to automatically evaluate the quality of summaries. The chapter, besides analyzing the most recent and used evaluation metrics, proposes an automatic summary evaluation method that shows high-level correlation with human judgments.

Section 2, "Social Networks and Web News Summarization," consists of 5 chapters and illustrates how summarization approaches can provide a succinct representation of relevant information related to social media data, events, and user/community interests.

Chapter 6, "Social Network Integration in Document Summarization," introduces the summarization approaches applied to social media data. Indeed, the advent of online social networking sites (e.g., Facebook, YouTube) has revolutionized the way people communicate and declared the social media as a primary source of knowledge about key events. This chapter overviews the most recent approaches to social media summarization and methods for update summarization, network activity summarization, event-based summarization, and opinion summarization. Moreover, a review of the existing evaluation metrics oriented to capturing the intrinsic quality of summaries and their usefulness to the human user is also provided.

Chapter 7, "Approaches to Large-Scale User Opinion Summarization for the Web," deals with the new frontiers of summarization research, where the payoffs are high, the datasets often huge, and the tasks very complex. The main example is the opinion summarization of large datasets that generally include high degrees of noise and little editorial structure. This chapter aims at defining the best practices to design an opinion summarization system by identifying three major aspects: (1) simple and modular techniques, (2) domain knowledge, and (3) evaluation metrics.

Chapter 8, "Novel Text Summarization Techniques for Contextual Advertising," introduces several novel summarization approaches to extracting summaries from Web pages in order to improve the quality of suggested ads. A comparison with state-of-the-art techniques and an assessment of whether the proposed techniques can be successfully applied to contextual advertising are presented. The proposed methods achieve good performance with respect to well-known text summarization techniques.

Chapter 9, "NewSum: 'N-Gram Graph'-Based Summarization in the Real World," describes the application of a summarization approach in a real use case scenario: multilingual news summarization. The proposed system, named NewSum, deals with the issues caused by the nature of real-world news by exploiting the n-gram graph representation to perform sentence selection and redundancy removal for the summaries. Clustering approaches are also used to detect events and topics. An open architecture for responsive summarization in a mobile setting is also presented. To understand the market applicability of the system, a pool of non-experts evaluated the quality of the summaries and the usefulness of the application in reading news.

Chapter 10, "New Formats and Interfaces for Multi-Document News Summarization and its Evaluation," presents news summarizers. In particular, it overviews automatic methods based on temporal text mining and graph-based methods and discusses the challenges associated with evaluation frameworks. Moreover, the graphical interfaces developed for the analysis and the search of summaries are analyzed in-depth, and, according to the results of case studies, the fundamental aspects in designing effective summarization and document search interfaces are identified.

Section 3, "Multilingual Summarization," consists of 3 chapters and covers the multilingual aspect of the summarization task. The state-of-the-art works addressing this issue are analyzed with a specific focus on the data representation and the text mining methods exploited during the summarization process.

Chapter 11, "Multilingual Summarization Approaches," overviews various state-of-the-art methods with special emphasis on multilingual summarizers. Indeed, in the age of electronic media, online information is available in several different languages other than English, and automatic summarization systems can be an indispensable solution to reduce the information overload and redundancy. The ap-

proaches presented are grouped based on their characteristics and the exploited document representation to highlight their main differences and the performance achieved in different scenarios.

Chapter 12, "Aspects of Multilingual News Summarization," discusses recent frameworks based on Latent Semantic Analysis (LSA) that showed good performance across many different languages. Starting from these methods, the authors show how domain-specific aspects can be used and a compression and paraphrasing method can be plugged in. A discussion of summarization evaluation in different languages is also presented, and two new approaches are introduced.

Chapter 13, "Language Independent Summarization Approaches," analyzes the language-dependent challenges and the most relevant works focused on language-independent algorithms. The advantages and disadvantages of the discussed methods dealing with multilingual collections are analyzed. Finally, new perspectives are presented to design language-independent systems.

Document summarization is an appealing research field that can provide useful tools for improving the accessibility to large volumes of data contained in document collections. Good summarization systems may also help in acquiring knowledge about a specific domain (e.g., biology) quickly and without redundancy and in understanding event causes (e.g., historical events) and many other aspects of human knowledge published over heterogeneous text resources. The objective of this work is to provide a clear and consolidated view of current summarization methods and their application to real scenarios. We seek to explore new methods to model text data, extract the relevant information according to the user interests and evaluate the quality and usefulness of the extracted summaries. We believe that a considerable research effort will be required in the future to improve the quality of these systems, for instance, in terms of computational load and readability of the summary. Different application domains might also benefit from summarization techniques and allow scientists to enhance their works. We also believe that multilingual approaches will be the next generation of summarization systems.

We are confident that professionals, researchers, and students in the fields of text mining, natural language processing, text summarization, and social network analysis will be able to use this book to learn more about the ways in which summarization techniques can prove useful in different environments.

Alessandro Fiori
Institute for Cancer Research and Treatment (IRCC), Italy