Preface

The interdisciplinary research area of Knowledge Discovery in Databases and Data Mining started forming in the late 80's in order to deal with the problem of automating data analysis, gathering the interest of researchers and practitioners from areas such as statistics, databases, and machine learning. Since then, the technologies for data sensing and storing have progressed enormously. Soon we will be able to record everything: personal multimedia, textual and other data, scientific structured data, business structured and unstructured data. Dealing with this information avalanche is still a grand challenge for data mining.

The ever-growing size of data being stored in today's information systems, inevitably leads to distributed database architectures, as data cannot fit in a single machine. Moreover, the offices or departments of many organizations and companies are scattered across a country or across the world, and central storage of data is often inefficient. In addition, globalization, business-to-business commerce and online collaboration between organizations rose lately the need for inter-organizational data mining. Advances in network technologies as well as the increasing use of networks, the Internet and distributed computing have contributed to the wide accessibility of distributed information sources that could be exploited for discovering interesting local and global knowledge. This led to a substantial amount of research work on mining distributed data sources and a new field of knowledge discovery under the title of distributed data mining.

Distributed data mining approaches cover a wide variety of learning algorithms that produce all kinds of knowledge, such as classification models, association rules and clusters. Such approaches have to deal with problems such as data heterogeneity, horizontal and/or vertical distribution, replication, asymmetry, model and results integration, comprehensibility and communication cost. Distributed data mining is a new field, which still has a lot of unresolved problems and open challenges. Some of the important current and future trends in this promising field we believe to be: location and resource aware algorithms, data mining on the world-wide web, grid-based data mining and data mining in mobile environments.

This special issue of the Journal of Information Sciences addresses the issue of knowledge discovery from distributed information sources. Despite its narrow focus, it attracted 15 submissions of papers. The selection of the final papers was a hard task due to the high quality of most of the submitted papers. After a rigorous review process four papers were selected for publication in this special issue.

The paper by Shyu et al. presents a methodology for discovering category clusters from distributed web directories. They consider distributed document category sets as information sources and merge some of the categories under certain distributed constraints. The resulting document category set offers better classification performance as exhibited by experiments with real-world web data.

The paper by Giannadakis et al. introduces InfoGrid, a data integration middleware engine, designed to operate under a Grid framework. It focuses on providing information access services and offers all users a query system which is able to retain the familiarity with their specific scientific applications while being diverse, flexible and open at the same time. The paper also presents how the InfoGrid architecture can be used to encounter the following important issues in Knowledge Discovery tasks: Providing contextual features for a data table to be used for analysis and finding relevant background knowledge for a user.

The paper by Pittie et al. considers the problem of detecting dependencies among data streams and presenting the results in a mobile data mining system. It gives an overview of their MobiMine system, identifies the technical challenges and offers solutions. It further discusses in detail the two important algorithmic techniques of correlation and conditional dependency rules that the system uses for dependency detection.

Finally, the paper by Scotney and McClean deals with the problem of uncertainty and imprecision that often accompanies information in distributed databases. Their approach uses the Dempster-Shafer theory of evidence to represent uncertainty and provides a mechanism for combining data and their uncertainty from distributed information sources based on aggregation of evidence. This mechanism can be used to resolve inconsistencies and provide information on data properties and patterns.

We believe that the selected papers touch upon several interesting aspects of distributed data mining and hope that this issue will prove to be an enjoyable and insightful reading. We would like to thank all the contributing authors and especially the reviewers for their support in producing this special issue. In alphabetical order, these were: David Cheung, Mohamed Elfeky, Moustafa Ghanem, Joydeep Ghosh, Robert Grossman, Yike Guo, Vasant Honavar, Hilol Kargupta, Sally McClean, Grigorios Tsoumakas and Mohammed Zaki.

Ioannis Vlahavas is an associate professor at the Department of Informatics at the Aristotle University of Thessaloniki. He received his Ph.D. degree in Logic Programming Systems from the same University in 1988. During the first half of 1997 he has been a visiting scholar at the Department of CS at Purdue University. He specializes in logic programming, knowledge based and AI systems and he has published over 90 papers, 8 book chapters and co-authored 3 books in these areas. He has been involved in more than 15 research projects, leading most of them. He was the chairman of the 2nd Hellenic Conference on AI. He is in head of the Logic Programming and Intelligent Systems (LPIS) Group which works on Knowledge-based systems, Data Mining and Planning methodologies for Intelligent Systems.