Executive Summary of the Dagstuhl Seminar 06201 on

Combinatorial and Algorithmic Foundations

# of Pattern and Association Discovery

by

Rudolf Ahlswede, Alberto Apostolico, and Vladimir I. Levenshtein

The focus of this seminar has been on the completely new scenario and on the wild paradigm shift that are forced by the recent progresses of ICT (information and communication technology). The new scenario is that data and information accumulate at a pace that makes it no longer fit for direct human inspection. The paradigm shift is that, in contrast to a primeval, persistent tenet of traditional information science and technology, the bottleneck in communication is no longer represented by the channel or medium but rather by the limited perceptual bandwidth of the final user: more and more often, the time and resources that need to be invested in order to gain access to information happens to be disproportionate to fruition time and value, thereby defying the very purpose of access. Consequently, the challenge of maximizing the throughput to the final user has taken up entirely new meanings. The implications brought about by such a dramatic change in perspectives have barely begun to be perceived. A science and engineering of discovery is developing to meet these challenges, which promises to revolutionize many facets of human activity beginning with the basic notions and practices of scientific investigation itself.

Above all, the problem of data overload looms ominously ahead in almost every field of our society. Databases in the Tera byte, even Peta byte range are now not uncommon. Our ability to analyze and understand massive datasets lags far behind our ability to gather and store the data with the ever advancing computer technology. Thus, as unprecedented volumes of information are amassed, disseminated and shared at an increasing pace in the emerging information infrastructures, the effective access to, and manipulation of information will depend no longer only on the efficiency with which information itself is structured, compressed, transmitted, stored and retrieved. A new generation of computational techniques and tools is required to support the extraction and the discovery of useful knowledge from the rapidly growing volumes of data. Raw data is rarely of direct benefit. Its true value is reflected by our ability to extract information useful for decision support or for exploration and understanding of the phenomena exhibited in the data source.

Huge amounts of scientific and social data are being produced and some have been made public in various databases or have been rendered commercially available. These data include experimental/observational data in Physics and Chemistry, DNA and amino acid sequences in Biology, Marketing data, financial data, etc. Thus the scope of data ranges from the microscopic world as to the global and cosmic world. Facing with these "data with hidden values", however, the current status of technology for discovering new scientific laws and

Dagstuhl Seminar Proceedings 06201
Combinatorial and Algorithmic Foundations of Pattern and Association Discovery
http://drops.dagstuhl.de/opus/volltexte/2006/792

knowledge useful for decision making is still immature. As said, a new era of challenges is opening with knowledge discovery technology in most areas in sciences and social activities. Our aim is to develop formal and practical methods for knowledge discovery from large compilations of data in various areas, and simultaneously, systematize the methods so far developed and applied in practical fields toward a creation of knowledge discovery paradigms. The task of analyzing data to extract useful information behind it is becoming more and more difficult because of the huge volume of data and limitations in computational resources.

At some core level in these endeavors, it comes natural to identify the need for novel techniques supporting the automated discovery of patterns and their associations or “rules” in disparate contexts and media. The techniques developed along these lines find ad hoc incarnations in diverse fields but also feature a distinctively unifying flavor. For instance, searching for identical or similar substrings in strings is of paramount interest to software development and maintenance, philology or plagiarism detection in the humanities, inference of common ancestries in molecular genetics, comparison of geological evolutions, stereo matching for robot vision, etc.. Checking the equivalence (e.g. identity up to a rotation) of circular strings finds use in determining the homology of organisms with circular genomes, comparing closed curves in computer vision, establishing the equivalence of polygons in computer graphics, etc. Finding repeated patterns, symmetries and cadences in strings is of interest to data compression, detection of recurrent events in symbolic dynamics, genome studies, intrusion detection in distributed computer systems, etc. The techniques for these problems have coalesced into an established core of Optimization, Pattern Matching, and other specialties of Algorithmics.

It is therefore a worthwhile effort to try and extract from the application areas crisp formulation of primitives, and study them in a coordinated fashion. Both theory and practice benefit from such an experience, as an increased degree of awareness and unification is fed back to both sides. This seminar thus concentrated on combinatorial and algorithmic techniques of pattern matching and pattern discovery that are regarded as the enabling machinery for such a revolution.