Chapter 1

Research Challenges for Data Mining in Science and Engineering

Jiawei Han and Jing Gao
University of Illinois at Urbana-Champaign

Abstract

With the rapid development of computer and information technology in the last several decades, an enormous amount of data in science and engineering has been and will continuously be generated in massive scale, either being stored in gigantic storage devices or flowing into and out of the system in the form of data streams. Moreover, such data has been made widely available, e.g., via the Internet. Such tremendous amount of data, in the order of tera- to peta-bytes, has fundamentally changed science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for new, data-intensive methods to conduct research in science and engineering.

In this paper, we discuss the research challenges in science and engineering, from the data mining perspective, with a focus on the following issues: (1) *information network analysis*, (2) *discovery, usage, and understanding of patterns and knowledge*, (3) *stream data mining*, (4) *mining moving object data, RFID data, and data from sensor networks*, (5) *spatiotemporal and multimedia data mining*, (6) *mining text, Web, and other unstructured data*, (7) *data cube-oriented multidimensional online analytical mining*, (8) *visual data mining*, and (9) *data mining by integration of sophisticated scientific and engineering domain knowledge.*

1.1 Introduction

It has been popularly recognized that the rapid development of computer and information technology in the last twenty years has fundamentally changed almost every field in science and engineering, transforming many disciplines from data-poor to increasingly data-rich, and calling for the development of new, data-intensive methods to conduct research in science and engineering. Thus the new terms like, *data science* or *data-intensive engineering*, can be used to best characterize the data-intensive nature of today's science and engineering.

Besides the further development of database methods to efficiently store and manage peta-bytes of data online, making these archives easily and safely accessible via the Internet and/or a computing grid, another essential task is to develop powerful data mining tools to analyze such data. Thus, there is no wonder that data mining has also stepped on to the center stage in science and engineering.

Data mining, as the confluence of multiple intertwined disciplines, including *statistics, machine learning, pattern recognition, database systems, information retrieval*, *World-Wide Web*, *visualization*, and *many application domains*, has made great progress in the past decade [HK06]. To ensure that the advances of data mining research and technology will effectively benefit the progress of science and engineering, it is important to examine the challenges on data mining posed in data-intensive science and engineering and explore how to further develop the technology to facilitate new discoveries and advances in science and engineering.

1.2 Major research challenges

In this section, we will examine several major challenges raised in science and engineering from the data mining perspective, and point out some promising research directions.

1.2.1 Information network analysis

With the development of Google and other effective web search engines, information network analysis has become an important research frontier, with broad applications, such as social network analysis, web community discovery, terrorist network mining, computer network analysis, and network intrusion detection. However, information network research should go beyond explicitly formed, homogeneous networks (*e.g.*, web page links, computer networks, and terrorist e-connection networks) and delve deeply into *implicitly formed*, *heterogeneous*, and *multidimensional* information networks. Science and engineering provide us with rich opportunities on exploration of networks in this direction.

There are a lot of massive natural, technical, social, and information networks in science and engineering applications, such as gene, protein, and microarray networks in biology; highway transportation networks in civil engineering; topic- or theme-author-publication-citation networks in library science; and wireless telecommunication networks among commanders, soldiers and supply lines in a battle field. In such information networks, each node or link in a network contains *valuable, multidimensional information*, such as textual contents, geographic information, traffic flow, and other properties. Moreover, such networks could be highly *dynamic, evolving, and inter-dependent*.

Traditional data mining algorithms such as classification, market basket analysis, and cluster analysis commonly attempt to find patterns in a dataset containing independent, identically distributed (IID) samples. One can think of this process as learning a model for the node attributes of a homogeneous graph while ignoring the links between the nodes. A key emerging challenge for data mining is tackling the problem of mining richly structured, heterogeneous datasets [GD05]. The domains often consist of a variety of object types; the objects can be linked in a variety of ways. Naively applying traditional statistical inference procedures, which assume that instances are independent, can lead to inappropriate conclusions about the data. In fact, object linkage is knowledge that should be exploited.

Although a single link in a network could be noisy, unreliable, and sometimes misleading, valuable knowledge can be mined reliably among a large number of links in a massive information network. Our recent studies on information networks show that the power of such links in massive information networks should not be underestimated. They can be used for predictive modeling across multiple relations [YHYY06], for user-guided clustering across multiple relations [YHY05], for effective link-based clustering [JW02, YHY06], for distinguishing different objects with identical names [YHY07a], and for solving the veracity problem, *i.e.*, finding reliable facts among multiple conflicting web information providers [YHY07b]. The power of such links should be thoroughly explored in many scientific domains, such as in protein network analysis in biology and in the analysis of networks of research publications in library science as well as in each science/engineering discipline.

The most well known link mining task is that of link-based object ranking (LBR), which is a primary focus of the link analysis community. The objective of LBR is to exploit the link structure of a graph to order or prioritize the set of objects within the graph. Since the introduction of the most notable approaches, PageRank [PBMW98] and HITS [Kle99], many variations have been developed to rank one type [CDG$^+$98, Hav02, RD02] or multiple types of objects in the graph [JW02, SQCF05]. Also, the link-based object classification (LBC) problem has been studied. The task is to predict the class label for each object. The discerning feature of LBC that makes it different from traditional classification is that in many cases, the labels of related objects tend to be correlated. The challenge is to design algorithms for collective classification that exploit such correlations and jointly infer the categorical values associated with the objects in the graph [CDI98]. Another link-related task is entity resolution, which involves identifying the set of objects in a domain. The goal of entity resolution is to determine which references in the data refer to the same real-world entity. Examples of this problem arise in databases (de-duplication, data integration) [ACG02, DHM05], natural language processing (co-reference resolution, object consolidation) [PMM$^+$02, BG06], personal information management, and other fields. Recently, there has been significant interest in the use of links for improved entity resolution. The central idea is to consider, in addition to the attributes of the references to be resolved, the other references to which these are linked. These links may be, for example, co-author links between author references in bibliographic data, hierarchical links between spatial references in geo-spatial data,

or co-occurrence links between name references in natural language documents. Besides utilizing links in data mining, we may wish to predict the existence of links based on attributes of the objects and other observed links in some problems. Examples include predicting links among actors in social networks, such as predicting friendships; predicting the participation of actors in events [OHS05], such as email, telephone calls and co-authorship; and predicting semantic relationships such as "advisor-of" based on web page links and content [CDF+00]. Most often, some links are observed, and one is attempting to predict unobserved links [GFKT01], or there is a temporal aspect.

Another important direction in information network analysis is to treat information networks as graphs and further develop graph mining methods [CH07]. Recent progress on graph mining and its associated structural pattern-based classification and clustering, graph and graph containment indexing, and similarity search will play an important role in information network analysis. An area of data mining that is related to link mining is the work on subgraph discovery. This work attempts to find interesting or commonly occurring subgraphs in a set of graphs. Discovery of these patterns may be the sole purpose of the systems, or the discovered patterns may be used for graph classification, whose goal is to categorize an entire graph as a positive or negative instance of a concept. One line of work attempts to find frequent subgraphs [KK01, YH07], and some other lines of work are on efficient subgraph generation and compression-based heuristic search [WM03, CH07]. Moreover, since information networks often form huge, multidimensional heterogeneous graphs, mining noisy, approximate, and heterogeneous subgraphs based on different applications for the construction of application-specific networks with sophisticated structures will help information network analysis substantially. Generative models for a range of graph and dependency types have been studied extensively in the social network analysis community [CSW05]. In recent years, significant attention has focused on studying the structural properties of networks [AC05], such as the World Wide Web, online social networks, communication networks, citation networks, and biological networks. Across these various networks, general patterns such as power law degree distributions, small graph diameters, and community structure are observed. These observations have motivated the search for general principles governing such networks [Cha05]. The use of the power law distribution of many information networks and the rules on density evolution of information networks will help reduce computational complexity and enhance the power of network analysis. Finally, the studies of link analysis, heterogeneous data integration, user-guided clustering, and user-based network construction will provide essential methodology for the in-depth study in this direction.

Many domains of interest today are best described as a network of interrelated heterogeneous objects. As future work, link mining may focus on the integration of link mining algorithms for a spectrum of knowledge discovery tasks. Furthermore, in many applications, the facts to be analyzed are dynamic and it is important to develop incremental link mining algorithms. Besides mining knowledge from links, objects and networks, we may wish to construct an information network based on both ontological and unstructured information.

1.2.2 Discovery, understanding, and usage of patterns and knowledge

Scientific and engineering applications often handle massive data of high dimensionality. The goal of pattern mining is to find itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. Pattern analysis can be a valuable tool for finding correlations, clusters, classification models, sequential and structural patterns, and outliers.

Frequent pattern mining has been a focused theme in data mining research for over a decade [HCXY07]. Abundant literature has been dedicated to this research, and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structural pattern mining, correlation mining, associative classification, and frequent-pattern-based clustering, as well as their broad applications.

The most focused and extensively studied topic in frequent pattern mining is perhaps scalable mining methods. There are also various proposals on reduction of such a huge set, including closed patterns, maximal patterns, approximate patterns, condensed pattern bases, representative patterns, clustered patterns, and discriminative frequent patterns. Recently, studies have proceeded to scalable methods for mining colossal patterns [ZYH+07] where the size of the patterns could be rather large so that the step-by-step growth using an Apriori-like approach does not work, and methods for pattern compression and extraction of high-quality top-k patterns [XCYH06]. Much research is still needed to substantially reduce the size of derived pattern sets, mine such patterns directly and efficiently, and enhance the quality of retained patterns.

Moreover, frequent pattern mining could help in other data mining tasks and many such pattern-based mining methods have been developed. Frequent patterns have been used for effective classification by association rule mining (such as [LHM98]), top-k rule generation for long patterns (such as [CTTX05]), and discriminative frequent pattern-based classification [WK05]. Recent studies show that better classification models could be constructed using discriminative frequent patterns and such patterns could be mined efficiently and directly from data [CYHH07, CYHY08]. Frequent patterns have also been used for clustering of high-dimensional biological data [WWYY02]. Therefore, frequent patterns can play an essential role in these major data mining tasks and the benefits should be exploited in depth.

We also need mechanisms for deep understanding and interpretation of patterns, e.g., semantic annotation for frequent patterns, and contextual analysis of frequent patterns. The main research work on pattern analysis has been focused on pattern composition (e.g., the set of items in item-set patterns) and frequency. A contextual analysis of frequent patterns over the structural information can help respond questions like "why this pattern is frequent?" [MXC+07]. The deep understanding of frequent patterns is essential to improve the interpretability and the usability of frequent patterns.

Besides studies on transaction datasets, much research has been done on effective sequential and structural pattern mining methods and the exploration of their applications [HCXY07, CH07]. Applications often raise new research issues and bring deep insight on the strength and weakness of an existing solution. Much work is needed to explore new applications of frequent pattern mining, for example, bioinformatics and software engineering.

The promotion of effective application of pattern analysis methods in scientific and engineering applications is an important task in data mining. Moreover, it is important to further develop efficient methods for mining long, approximate, compressed, and sophisticated patterns for advanced applications, such as mining biological sequences and networks and mining patterns related to scientific and engineering processes. Furthermore, the exploration of mined patterns for classification, clustering, correlation analysis, and pattern understanding will still be interesting topics in research.

1.2.3 Stream data mining

Stream data refers to the data that flows into and out of the system like streams. Stream data is usually in vast volume, changing dynamically, possibly infinite, and containing multi-dimensional features. Typical examples of such data include audio and video recording of scientific and engineering processes, computer network information flow, web click streams, and satellite data flow. Such data cannot be handled by traditional database systems, and moreover, most systems may only be able to read a data stream once in sequential order. This poses great challenges on effective mining of stream data [BBD+02, Agg06].

First, the techniques to summarize the whole or part of the data streams are studied, which is the basis for stream data mining. Such techniques include sampling [DH01], load shedding [TcZ+03] and sketching techniques [Mut03], synopsis data structures [GKMS01], stream cubing [CDH+02], and clustering [AHWY03]. Progress has been made on efficient methods for mining frequent patterns in data streams [MM02], multidimensional analysis of stream data (such as construction of stream cubes) [CDH+02], stream data classification [AHWY04], stream clustering [AHWY03], stream outlier analysis, rare event detection [GFHY07], and so on. The general philosophy is to develop single-scan algorithms to collect information about stream data in tilted time windows, exploring micro-clustering, limited aggregation, and approximation.

The focus of stream pattern analysis is to approximate the frequency counts for infinite stream data. Algorithms have been developed to count frequency using tilted windows [GHPY02] based on the fact that users are more interested in the most recent transactions; approximate frequency counting based on previous historical data to calculate the frequent patterns incrementally [MM02] and track the most frequent k items in the continuously arriving data [CM03].

Initial studies on stream clustering concentrated on extending K-means and K-median algorithms to stream environment [GMM+03]. The main idea behind the developed algorithms is that the cluster centers and weights are updated after examining one transaction or a batch of transactions, whereas the constraints on memory and time complexity are satisfied by limiting the number of centers. Later, [AHWY03] proposes to divide the clustering process into online microclustering process, which stores summarized statistics about the data streams, and the offline one, which performs macro-clustering on the summarized data according to a number of user preferences such as the time frame and the number of clusters. Projected clustering can also be perform for high dimensional

data streams [AHWY04].

The focus of stream classification of data streams is first on how to efficiently update the classification model when data continuously flow in. VFDT [DH00] is a representative method in this field where a incremental decision tree is built based on Hoeffding trees. Later, the concept drift problem in data stream classification has been recognized, which refers to the unknown changes of the distribution underlying data streams. Many algorithms have been developed to prevent deterioration in prediction accuracy of the model [HSD01, KM05], by carefully selecting training examples that represent the true concept [Fan04] or combining multiple models to reduce variance in prediction [WFYH03, GFH07]. For skewed distribution of stream data, it is recommended to explore biased selective sampling and robust ensemble methods in model construction [GFHY07].

Stream data is often encountered in science and engineering applications. It is important to explore stream data mining in such applications and develop application-specific methods, *e.g.*, real-time anomaly detection in computer network analysis, in electric power grid supervision, in weather modeling, in engineering and security surveillance, and other stream data applications.

1.2.4 Mining moving object data, RFID data, and data from sensor networks

With the popularity of sensor networks, GPS, cellular phones, other mobile devices, and RFID technology, tremendous amount of moving object data has been collected, calling for effective analysis. This is especially true in many scientific, engineering, business and homeland security applications.

Sensor networks are finding increasing number of applications in many domains, including battle fields, smart buildings, and even the human body. Most sensor networks consist of a collection of light-weight (possibly mobile) sensors connected via wireless links to each other or to a more powerful gateway node that is in turn connected with an external network through either wired or wireless connections. Sensor nodes usually communicate in a peer-to-peer architecture over an asynchronous network. In many applications, sensors are deployed in hostile and difficult to access locations with constraints on weight, power supply, and cost. Moreover, sensors must process a continuous (possibly fast) stream of data. Data mining in wireless sensor networks (WSNs) is a challenging area, as algorithms need to work in extremely demanding and constrained environment of sensor networks (such as limited energy, storage, computational power, and bandwidth). WSNs also require highly decentralized algorithms.

Development of algorithms that take into consideration the characteristics of sensor networks, such as energy and computation constraints, network dynamics, and faults, constitute an area of current research. Some work has been done in developing localized, collaborative, distributed and self-configuration mechanisms in sensor networks. In designing algorithms for sensor networks, it is imperative to keep in mind that power consumption has to be minimized. Even gathering the distributed sensor data in a single site could be expensive in terms of battery power consumed, some attempts have been made towards making the data collection task energy efficient and balance the energy-quality trade-offs. Clustering the nodes of the sensor networks is an important optimization problem. Nodes that are clustered together can easily communicate with each other, which can be applied to energy optimization and developing optimal algorithms for clustering sensor nodes. Other works in this field include identification of rare events or anomalies, finding frequent itemsets, and data preprocessing in sensor networks.

Recent years have witnessed and enormous increase in moving object data from RFID records in supply chain operations, toll and road sensor readings from vehicles on road networks, or even cell phone usage from different geographic regions. These movement data, including RFID data, object trajectories, anonymous aggregate data such as the one generated by many road sensors, contain rich information. Effective management of such data is a major challenge facing society today, with important implications into business optimization, city planning, privacy, and national security. Interesting research has been conducted on warehousing RFID data sets [GHLK06], which could handle moving object data sets by significantly compressing such data, and proposing a new aggregation mechanism that preserves their path structures. Mining moving objects is a challenging problem due to the massive size of the data, and its spatiotemporal characteristics. The methods developed along this line include FlowGraph [GHL06b], which is a probabilistic model that captures the main trends and exceptions in moving object data, and FlowCube [GHL06a], which is a multi-dimensional extension of the FlowGraph and an adaptive fastest path algorithm [GHL$^+$07] that computes routes based on driving patterns present in the data. RFID systems are known to generate noisy data so data cleaning is an essential task for the correct interpretation and analysis of moving object data, especially when it is collected from RFID applications and thus demans for cost-effective cleaning methods (such as [GHS07]). One important application with moving objects is automated

identification of suspicious movements. A framework for detecting anomalies [LHKG07] is proposed to express object trajectories using discrete pattern fragments, extract features to form a hierarchical feature space and learn effective classification rules at multiple levels of granularity. Another line of work on outlier detection in trajectories focuses on detecting outlying sub-trajectories [LHL08] based on partition-and-detect framework, which partitions a trajectory into a set of line segments, and then, detects outlying line segments for trajectory outliers. The problem of clustering trajectory data [LHW07] is also studied where common sub-trajectories are discovered using the minimum description length (MDL) principle.

Overall, this is still a young field with many research issues to be explored on mining moving object data, RFID data, and data from sensor networks. For example, how to explore correlation and regularity to clean noisy sensor network and RFID data, how to integrate and construct data warehouses for such data, how to perform scalable mining for peta-byte RFID data, how to find strange moving objects, how to classify multidimensional trajectory data, and so on. With time, location, moving direction, speed, as well as multidimensional semantics of moving object data, likely multi-dimensional data mining will play an essential role in this study.

1.2.5 Spatial, temporal, spatiotemporal, and multimedia data mining

Scientific and engineering data is usually related to space, time, and in multimedia modes (*e.g.*, containing color, image, audio, and video). With the popularity of digital photos, audio DVDs, videos, YouTube, web-based map services, weather services, satellite images, digital earth, and many other forms of multimedia, spatial, and spatiotemporal data, mining spatial, temporal, spatiotemporal, and multimedia data will become increasingly popular, with far-reaching implications [MH01, SC03]. For example, mining satellite images may help detect forest fire, find unusual phenomena on earth, predict hurricane landing site, discover weather patterns, and outline global warming trends.

Spatial data mining is the process of discovering interesting and previously unknown, but potentially useful patterns from large spatial data sets [SZHV04]. Extracting interesting and useful patterns from spatial data sets is more difficult than extracting the corresponding patterns from traditional numeric and categorical data due to the complexity of spatial data types, spatial relationships, and spatial autocorrelation. Interesting research topics in this field include prediction of events at particular geographic locations, detecting spatial outliers whose no-spatial attributes are extreme relative to its neighbors, finding co-location patterns where instances containing the patterns often located in close geographic proximity, and grouping a set of spatial objects into clusters. Future research is needed to compare the difference and similarity between classical data mining and spatial data mining techniques, model semantically rich spatial properties other than neighborhood relationships, design effective statistical methods to interpret the mined spatial patterns, investigate proper measures for location prediction to improve spatial accuracy and facilitate visualization of spatial relationships by representing both spatial and non-spatial features. The problems of incorporating domain knowledge into mining when data is scarce and integrating data collection with mining are worth studying in spatial data mining, and both theoretical analyses toward general studies of spatial phenomena and empirical model designs targeted for specific applications represent the trends for future research.

Time series data [SZ04], which represent sequences of recorded values, appear naturally in almost all fields of applications including physics, finance, medicine and music. People have tried to obtain insights into the mechanism that generates and the time series and use the valuable information contained in time series to predict future values. In the last decade there has been an explosion of interest in mining time series data. Literally hundreds of papers have introduced new algorithms to preprocess, index, classify, cluster, and identify patterns or novelties from time series. As future work, the research on time series should consider mining multiple time series of the same type or of different types, incorporating domain knowledge into time series mining and facilitate real time time series mining in some applications.

For applications involving multimedia data, we need tools for discovering relationships between objects or segments within multimedia document components, such as classifying images based on their contents, extracting patterns in sound, categorizing speech and music, and recognizing and tracking objects in video streams. In general, the multimedia files from a database must be first preprocessed to improve their quality. Subsequently, these multimedia files undergo various transformations and features extraction to generate important features from the multimedia files. With the generated features, mining can be carried out using data mining techniques to discover significant patterns. These resulting patterns are then evaluated and interpreted in order to obtain

the final applications knowledge. Numerous methodologies have been developed and many applications have been investigated, including organizing multimedia data indexing and retrieval, extracting representative features from raw multimedia data before the mining process and integrateing features obtained from multiple modalities. For example, the MPEG-7 standard provides a good representative set of features. Automatic annotation, also referred to as concept mining, is one of the main tasks in multimedia mining. The methods developed for this task include supervised learning, unsupervised learning and contexts-based approaches. In supervised learning, based on the annotated concepts for each multimedia document, the unclassified files are automatically categorized. In unsupervised learning, multimedia files are clustered and annotators assign key words to each cluster, which could be used to extract rules for annotating future documents [SS03]. The third approach tries to mine concepts by looking at the contextual information [SW05], such as the text associated with images, to derive semantic concepts. Another important topic, detection of interesting or unusual events, has received considerable interest in multimedia research. In the future, multimedia mining will be continuously receiving attention, especially for its application in online video sharing, security surveillance monitoring and effective image retrieval.

Spatiotemporal data mining [RHS01] is an emerging research area that is dedicated to the development of novel algorithms and computational techniques for the successful analysis of large spatiotemporal databases and the disclosure of interesting knowledge in spatiotemporal data. Much work has been done to modify the data mining techniques so that they can, to the largest extent, exploit the rich spatiotemporal relationships/patterns embedded in the datasets. Spatiotemporal data mining tasks and techniques can be roughly classified into indexing and searching, pattern analysis, clustering, compression, and outlier detection.

Both the temporal and spatial dimensions could add substantial complexity to data mining tasks. First, the spatial and temporal relationships are information bearing and therefore need to be considered in data mining. Some spatial and temporal relationships are implicitly defined, and must be extracted from the data. Such extraction introduces some degree of fuzziness and/or uncertainty that may have an impact on the results of the data mining process. Second, working at the level of stored data is often undesirable, and thus complex transformations are required to describe the units of analysis at higher conceptual levels. Third, interesting patterns are more likely to be discovered at the lowest resolution/granularity level, but large support is more likely to exist at higher levels. Finally, how to express domain independent knowledge and how to integrate spatiotemporal reasoning mechanisms in data mining systems are still open problems.

Research in this domain needs the confluence of multiple disciplines including image processing, pattern recognition, geographic information systems, parallel processing, and statistical data analysis. Automatic categorization of images and videos, classification of spatiotemporal data, finding frequent/sequential patterns and outliers, spatial collocation analysis, and many other tasks have been studied popularly. With the mounting of such data, the development of scalable analysis methods and new data mining functions will be an important research frontier for years to come.

1.2.6 Mining text, Web, and other unstructured data

Web is the common place for scientists and engineers to publish their data, share their observations and experiences, and exchange their ideas. There is a tremendous amount of scientific and engineering data on the web. For example, in biology and bioinformatics research, there are GenBank, ProteinBank, GO, PubMed, and many other biological or biomedical information repositories available on the Web. Therefore, the Web has become the ultimate information access and processing platform, housing not only billions of link-accessed "pages", containing textual data, multimedia data, and linkages, on the surface Web, but also query-accessible "databases" on the deep Web. With the advent of Web 2.0, there is an increasing amount of dynamic "workflow" emerging. With its penetrating deeply into our daily life and evolving into unlimited dynamic applications, the Web is central in our information infrastructure. Its virtually unlimited scope and scale render immense opportunities for data mining.

Text mining and information extraction [Ber03] have been applied not only to Web mining but also to the analysis of other kinds of semi-structured and unstructured information, such as digital libraries, biological information systems, research literature analysis systems, computer-aided design and instruction, and office automation systems. Technologies in the text-mining process include information extraction, topic tracking, summarization, categorization, clustering, and concept linkage. Information extraction [Cha01] represents a starting point for computers analyzing unstructured text and identifying key phrases and relationships within text. It does it by looking for predefined sequences in the text, a process called pattern matching. A topic-tracking system [All02]

keeps user profiles and, based on the documents a user views, predicts other documents of interest to the user. Text summarization [ER04] helps users figure out whether a lengthy document meets their needs and is worth reading. With large-volume texts, text-summarization software processes and summarizes the document in almost no time. The key to summarization is reducing the length and detail of a document while retaining its main points and overall meaning. Categorization involves identifying the main themes of a document [WIZD04]. When categorizing particular documents, a computer program often treats them as a "bag of words". The program does not attempt to process the actual information as information extraction does. Rather, categorization counts only words that appear and, from the counts, identifies the main topics covered in the document. Clustering is a technique used to group similar documents [DM01], but it differs from categorization in that documents are clustered on the fly instead of through predefined topics. Documents can also appear in multiple subtopics, ensuring that useful documents are not omitted from the search results. Some topic-modeling techniques [MZ05] connect related documents by identifying their shared concepts, helping users find information they perhaps wouldn't have found through traditional search methods. It promotes browsing for information rather than searching for it.

Web Mining [Liu06, Cha02, KB00] is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World-Wide Web. There are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is an automatic process that goes beyond keyword extraction [QD07] to discover useful information from the content of a web page. The type of the web content may consist of text, image, audio or video data in the web. The text content is the most widely researched area. The technologies that are normally used in web content mining are natural language processing, information retrieval, and text mining. The strategies include directly mining the content of documents and improving on the content search of other tools like search engines. Web Structure Mining [FLGC02] can help reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. Counters of hyperlinks, in and out documents, retrace the structure of the web artifacts summarized. Finally, Web Usage Mining [FL05] can help understand the user behavior and the web structure by analyzing the web access logs of different web sites. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking, which analyzes the web logs to understand access patterns and trends to construct better structure and grouping of resource providers, and Customized Usage Tracking, which analyzes individual trends to customize web sites to users.

There are lots of research issues in this domain [Cha02, Liu06], which takes collaborative efforts of multiple disciplines, including information retrieval, databases, data mining, natural language processing, and machine learning. For many scientific and engineering applications, the data is somewhat structured and semi-structured, with designated fields for text and multimedia data. Thus it is possible to mine and build relatively structured web repositories. Some promising research topics include heterogeneous information integration, information extraction, personalized information agents, application-specific partial Web construction and mining, in-depth Web semantics analysis, development of scientific and engineering domain-specific semantic Webs, and turning Web into relatively structured information-base.

1.2.7 Data cube-oriented multidimensional online analytical mining

Scientific and engineering datasets are usually high-dimensional in nature. Viewing and mining data in multidimensional space will substantially increase the power and flexibility of data analysis. Data cube computation and OLAP (online analytical processing) technologies developed in data warehouse have substantially increased the power of multidimensional analysis of large datasets.

Some researchers began to investigate how to conduct traditional data mining and statistical analysis in the multi-dimensional manner efficiently. For example, regression cube [CDH$^+$06] is designed to support efficient computation of the statistical models. In this framework, each cell can be compressed into an auxiliary matrix with a size independent of the number of tuples and then the statistical measures for any data cell can be computed from the compressed data of the lower-level cells without accessing the raw data. In a prediction cube [CCLR05], each cell contains a value that summarizes a predictive model trained on the data corresponding to that cell and characterizes its decision behavior or predictiveness. The authors further show that such cubes can be efficiently computed by exploiting the idea of model decomposition. In [LH07], the issues of anomaly detection in multi-

dimensional time-series data are examined. A time-series data cube is proposed to capture the multi-dimensional space formed by the attribute structure and facilitate the detection of anomalies based on expected values derived from higher level, more general time-series. Moreover, an efficient search algorithm is proposed to iteratively select subspaces in the original high-dimensional space and detect anomalies within each one. Recent study on sampling cubes [LHY+08] discuss about the desirability of OLAP over sampling data, which may not represent the full data in the population. The proposed sampling cube framework could efficiently calculate confidence intervals for any multidimensional query and uses the OLAP structure to group similar segments to increase sampling size when needed. Further, to handle high dimensional data, a Sampling Cube Shell method is proposed to effectively reduce the storage requirement while still preserving query result quality. Such multi-dimensional, especially high-dimensional, analysis tools will ensure data can be analyzed in hierarchical, multidimensional structures efficiently and flexibly at user's finger tips. This leads to the integration of online analytical processing with data mining, *i.e.*, OLAP mining. Some efforts have been devoted along this direction, but grand challenge still exits when one needs to explore the large space of choices to find interesting patterns and trends [RC07].

We believe that OLAP mining will substantially enhance the power and flexibility of data analysis and lead to the construction of easy-to-use tools for the analysis of massive data with hierarchical structures in multidimensional space. It is a promising research field for developing effective tools and scalable methods for exploratory-based scientific and engineering data mining.

1.2.8 Visual data mining

A picture is worth a thousand words. There have been numerous data visualization tools for visualizing various kinds of data sets in massive amount and of multidimensional space [Tuf01]. Besides popular bar charts, pie charts, curves, histograms, quantile plots, quantitle-quantile plots, boxplots, scatter plots, there are also many visualization tools using geometric (*e.g.*, dimension stacking, parallel coordinates), hierarchical (*e.g.*, treemap), and icon-based (*e.g.*, Chernoff faces and stick figures) techniques. Moreover, there are methods for visualizing sequences, time-series data, phylogenetic trees, graphs, networks, web, as well as various kinds of patterns and knowledge (*e.g.*, decision-trees, association rules, clusters and outliers) [FGW01]. There are also visual data mining tools that may facilitate interactive mining based on user's judgement of intermediate data mining results [AEEK99]. Recently, we have developed a DataScope system that maps relational data into 2-D maps so that multidimensional relational data can be browsed in Google map's way [WLX+07].

Most data analysts use visualization as part of a process sandwich strategy of interleaving mining and visualization to reach a goal, an approach commonly identified in many research works on applications and techniques for visual data mining [dOL03]. Usually, the analytical mining techniques themselves do not rely on visualization. Most of the papers describing visual data mining approaches and applications found in the literature fall into two categories: either they use visual data exploration systems or techniques to support a knowledge extraction goal or a specific mining task, or they use visualization to display the results of a mining algorithm, such as a clustering process or a classifier, and thus enhance user comprehension of the results. A classification of information visualization and visual data mining techniques is proposed in [Kei02], which is based on the data type to be visualized, the visualization technique, and the interaction and distortion technique. Mining tasks usually demand techniques capable of handling large amounts of multidimensional data, often in the format of Data Tables or relational databases. Parallel coordinates and scatter plots are much exploited in this context. Also, interaction mechanisms for filtering, querying, and selecting data are typically required for handling larger data sets [Ins97]. Another typical use of visualization in mining resides in visually conveying the results of a mining task, such as clustering or classification, to enhance user interpretation. One such example is given by the BLOB clustering algorithm [GSF97], which uses implicit surfaces for visualizing data clusters. But rather than using visual data exploration and analytical mining algorithms as separate tools, a stronger data mining strategy would be to tightly couple the visualizations and analytical processes into one data mining tool. Many mining techniques involve different mathematical steps that require user intervention. Some of these can be quite complex and visualization can support the decision processes involved in making such interventions. From this viewpoint, a visual data mining technique is not just a visualization technique being applied to exploit data in some phases of an analytical mining process, but a data mining algorithm in which visualization plays a major role.

We believe that visual data mining is appealing to scientists and engineers because they often have good understanding of their data, can use their knowledge to interpret their data and patterns with the help of visual-

ization tools, and interact with the system for deeper and more effective mining. Tools should be developed for mapping data and knowledge into appealing and easy-to-understand visual forms, and for interactive browsing, drilling, scrolling, and zooming data and patterns to facilitate user exploration. Finally, for visualization of large amount of data, parallel processing and high-performance visualization tools should be investigated to ensure high performance and fast response.

1.2.9 Domain-specific data mining: Data mining by integration of sophisticated scientific and engineering domain knowledge

Besides general data mining methods and tools for science and engineering, each scientific or engineering discipline has its own data sets and special mining requirements, some could be rather different from the general ones. Therefore, in-depth investigation of each problem domain and development of dedicated analysis tools are essential to the success of data mining in this domain. Here we examine two problem domains: biology and software engineering.

Biological data mining

The fast progress of biomedical and bioinformatics research has led to the accumulation and publication (on the web) of vast amount of biological and bioinformatics data. However, the analysis of such data poses much greater challenges than traditional data analysis methods [BHLY04]. For example, genes and proteins are gigantic in size (*e.g.*, a DNA sequence could be in billions of base pairs), very sophisticated in function, and the patterns of their interactions are largely unknown. Thus it is a fertile field to develop sophisticated data mining methods for in-depth bioinformatics research. We believe substantial research is badly needed to produce powerful mining tools in many biological and bioinformatics subfields, including comparative genomics, evolution and phylogeny, biological data cleaning and integration, biological sequence analysis, biological network analysis, biological image analysis, biological literature analysis (*e.g.*, PubMed), and systems biology. From this point view, data mining is still very young with respect to biology and bioinformatics applications. Substantial research should be conducted to cover the vast spectrum of data analysis tasks.

Data mining for software engineering

Software program executions potentially (*e.g.*, when program execution traces are turned on) generate huge amounts of data. However, such data sets are rather different from the datasets generated from the nature or collected from video cameras since they represent the executions of program logics coded by human programmers. It is important to mine such data to monitor program execution status, improve system performance, isolate software bugs, detect software plagiarism, analyze programming system faults, and recognize system malfunctions.

Data mining for software engineering can be partitioned into static analysis and dynamic/stream analysis, based on whether the system can collect traces beforehand for post-analysis or it must react at real time to handle online data. Different methods have been developed in this domain by integration and extension of the methods developed in machine learning, data mining, pattern recognition, and statistics. For example, statistical analysis such as hypothesis testing) approach [LFY+06] can be performed on program execution traces to isolate the locations of bugs which distinguish program success runs from failing runs. Despite of its limited success, it is still a rich domain for data miners to research and further develop sophisticated, scalable, and real-time data mining methods.

1.3 Conclusions

Science and engineering are fertile lands for data mining. In the last two decades, science and engineering have evolved to a stage that gigantic amounts of data are constantly being generates and collected, and data mining and knowledge discovery becomes the essential scientific discovery process. We have proceeded to the era of *data science* and *data engineering*.

In this paper, we have examined a few important research challenges in science and engineering data mining. There are still several interesting research issues not covered in this short abstract. One such issue is the development of *invisible data mining* functionality for science and engineering which builds data mining functions as an invisible process in the system (*e.g.*, rank the results based on the relevance and some sophisticated, preprocessed evaluation functions) so that users may not even sense that data mining has been performed beforehand or is

being performed and their browsing and mouse clicking are simply using the results of or further exploring of data mining. Another research issue is *privacy-preserving data mining* that aims to performing effective data mining without disclosure of private or sensitive information to outsiders. Finally, *knowledge-guided intelligent human computer interaction* based on the knowledge extracted from data could be another interesting issue for future research.

1.4 Acknowledgments

The work was supported in part by the U.S. National Science Foundation NSF IIS-05-13678 and BDI-05-15813. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

Bibliography

[AC05] Edoardo M. Airoldi and Kathleen M. Carley. Sampling algorithms for pure network topologies: a study on the stability and the separability of metric embeddings. *SIGKDD Explor. Newsl.*, 7(2):13–22, 2005.

[ACG02] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 586–597, Hong Kong, China, Aug. 2002.

[AEEK99] M. Ankerst, C. Elsen, M. Ester, and H.-P. Kriegel. Visual classification: An interactive approach to decision tree construction. In *Proc. 1999 Int. Conf. Knowledge Discovery and Data Mining (KDD'99)*, pages 392–396, San Diego, CA, Aug. 1999.

[Agg06] C. C. Aggarwal. *Data Streams: Models and Algorithms*. Kluwer Academic, 2006.

[AHWY03] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for clustering evolving data streams. In *Proc. 2003 Int. Conf. Very Large Data Bases (VLDB'03)*, pages 81–92, Berlin, Germany, Sept. 2003.

[AHWY04] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu. A framework for projected clustering of high dimensional data streams. In *Proc. 2004 Int. Conf. Very Large Data Bases (VLDB'04)*, pages 852–863, Toronto, Canada, Aug. 2004.

[All02] James Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[BBD+02] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In *Proc. 2002 ACM Symp. Principles of Database Systems (PODS'02)*, pages 1–16, Madison, WI, June 2002.

[Ber03] M. W. Berry. *Survey of Text Mining: Clustering, Classification, and Retrieval*. Springer, 2003.

[BG06] I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. In *Proc. 2006 SIAM Int. Conf. Data Mining (SDM'06)*, Bethesda, MD, April 2006.

[BHLY04] P. Bajcsy, J. Han, L. Liu, and J. Yang. Survey of bio-data analysis from data mining perspective. In Jason T. L. Wang, Mohammed J. Zaki, Hannu T. T. Toivonen, and Dennis Shasha, editors, *Data Mining in Bioinformatics*, pages 9–39. Springer Verlag, 2004.

[CCLR05] B.-C. Chen, L. Chen, Y. Lin, and R. Ramakrishnan. Prediction cubes. In *Proc. 2005 Int. Conf. Very Large Data Bases (VLDB'05)*, pages 982–993, Trondheim, Norway, Aug. 2005.

[CDF+00] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118:69–113, 2000.

[CDG+98] S. Chakrabarti, B. E. Dom, D. Gibson, J. M. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. 7th Int. World Wide Web Conf. (WWW'98)*, pages 65–74, Brisbane, Australia, 1998.

[CDH+02] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multi-dimensional regression analysis of time-series data streams. In *Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 323–334, Hong Kong, China, Aug. 2002.

[CDH+06] Y. Chen, G. Dong, J. Han, J. Pei, B. W. Wah, and J. Wang. Regression cubes with lossless compression and aggregation. *IEEE Trans. Knowledge and Data Engineering*, 18:1585–1599, 2006.

[CDI98] S. Chakrabarti, B. E. Dom, and P. Indyk. Enhanced hypertext classification using hyper-links. In *Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98)*, pages 307–318, Seattle, WA, June 1998.

[CH07] D. J. Cook and L. B. Holder. *Mining Graph Data*. John Wiley & Sons, 2007.

[Cha01] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *Proc. 2001 Int. World Wide Web Conf. (WWW'01)*, pages 211–220, Hong Kong, China, May 2001.

[Cha02] S. Chakrabarti. *Mining the Web: Statistical Analysis of Hypertex and Semi-Structured Data*. Morgan Kaufmann, 2002.

[Cha05] Deepayan Chakrabarti. *Tools for large graph mining*. PhD thesis, Pittsburgh, PA, USA, 2005. Chair-Christos Faloutsos.

[CM03] Graham Cormode and S. Muthukrishnan. What's hot and what's not: tracking most frequent items dynamically. In *PODS '03: Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 296–306, New York, NY, USA, 2003. ACM.

[CSW05] P. J. Carrington, J. Scott, and S. Wasserman. Models and methods in social network analysis. Cambridge University Press, 2005.

[CTTX05] G. Cong, K.-Lee Tan, A. K. H. Tung, and X. Xu. Mining top-k covering rule groups for gene expression data. In *Proc. 2005 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'05)*, pages 670–681, Baltimore, MD, June 2005.

[CYHH07] H. Cheng, X. Yan, J. Han, and C.-W. Hsu. Discriminative frequent pattern analysis for effective classification. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.

[CYHY08] H. Cheng, X. Yan, J. Han, and P. S. Yu. Direct discriminative pattern mining for effective classification. In *Proc. 2008 Int. Conf. Data Engineering (ICDE'08)*, Cancun, Mexico, April 2008.

[DH00] P. Domingos and G. Hulten. Mining high-speed data streams. In *Proc. 2000 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'00)*, pages 71–80, Boston, MA, Aug. 2000.

[DH01] Pedro Domingos and Geoff Hulten. A general method for scaling up machine learning algorithms and its application to clustering. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 106–113, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[DHM05] Xin Dong, Alon Halevy, and Jayant Madhavan. Reference reconciliation in complex information spaces. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 85–96, New York, NY, USA, 2005. ACM.

[DM01] Inderjit S. Dhillon and Dharmendra S. Modha. Concept decompositions for large sparse text data using clustering. *Mach. Learn.*, 42(1-2):143–175, 2001.

[dOL03] M.C. Ferreira de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):378–394, 2003.

[ER04] Gunes Erkan and Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artificial Intelligence Research*, 22:457–479, 2004.

[Fan04] Wei Fan. Systematic data selection to mine concept-drifting data streams. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 128–137, New York, NY, USA, 2004. ACM.

[FGW01] U. Fayyad, G. Grinstein, and A. Wierse. *Information Visualization in Data Mining and Knowledge Discovery*. Morgan Kaufmann, 2001.

[FL05] Federico Michele Facca and Pier Luca Lanzi. Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng.*, 53(3):225–241, 2005.

[FLGC02] G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35:66–71, 2002.

[GD05] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explorations*, 7:3 – 12, 2005.

[GFH07] J. Gao, W. Fan, and J. Han. On appropriate assumptions to mine data streams: Analysis and practice. In *Proc. 2007 Int. Conf. Data Mining (ICDM'07)*, Omaha, NE, Oct. 2007.

[GFHY07] J. Gao, W. Fan, J. Han, and P. S. Yu. A general framework for mining concept-drifting data streams with skewed distributions. In *Proc. 2007 SIAM Int. Conf. Data Mining (SDM'07)*, Minneapolis, MN, April 2007.

[GFKT01] L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of relational structure. In *Proc. 2001 Int. Conf. Machine Learning (ICML'01)*, pages 170–177, Williamstown, MA, 2001.

[GHL06a] H. Gonzalez, J. Han, and X. Li. Flowcube: Constructuing RFID flowcubes for multi-dimensional analysis of commodity flows. In *Proc. 2006 Int. Conf. Very Large Data Bases (VLDB'06)*, pages 834–845, Seoul, Korea, Sept. 2006.

[GHL06b] H. Gonzalez, J. Han, and X. Li. Mining compressed commodity workflows from massive rfid data sets. In *Proc. 2006 Int. Conf. Information and Knowledge Management (CIKM'06)*, Arlington, VA, Nov. 2006.

[GHL+07] H. Gonzalez, J. Han, X. Li, M. Myslinska, and J. P. Sondag. Adaptive fastest path computation on a road network: A traffic mining approach. In *Proc. 2007 Int. Conf. Very Large Data Bases (VLDB'07)*, Vienna, Austria, Sept. 2007.

[GHLK06] H. Gonzalez, J. Han, X. Li, and D. Klabjan. Warehousing and analysis of massive RFID data sets. In *Proc. 2006 Int. Conf. Data Engineering (ICDE'06)*, page 83, Atlanta, Georgia, April 2006.

[GHPY02] C. Giannella, J. Han, J. Pei, and P. S. Yu. FP-stream: Lazy mining of frequent patterns in data streams. In *Technical Report, Univ. of Illinois*, Oct. 2002.

[GHS07] H. Gonzalez, J. Han, and X. Shen. Cost-conscious cleaning of massive rfid data sets. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.

[GKMS01] A. C. Gilbert, Y. Kotidis, S. Muthukrishnan, and M. Strauss. Surfing wavelets on streams: One-pass summaries for approximate aggregate queries. In *Proc. 2001 Int. Conf. on Very Large Data Bases (VLDB'01)*, pages 79–88, Rome, Italy, Sept. 2001.

[GMM+03] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. *IEEE Trans. Knowledge and Data Engineering*, 15:515–528, 2003.

[GSF97] M. H. Gross, T. C. Sprenger, and J. Finger. Visualizing informationon a sphere. In *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*, page 11, 1997.

[Hav02] Taher H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM.

[HCXY07] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 15:55–86, 2007.

[HK06] J. Han and M. Kamber. *Data Mining: Concepts and Techniques* (2nd ed.). Morgan Kaufmann, 2006.

[HSD01] G. Hulten, L. Spencer, and P. Domingos. Mining time-changing data streams. In *Proc. 2001 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'01)*, San Fransisco, CA, Aug. 2001.

[Ins97] A. Inselberg. Multidimensional detective. In *Proceedings of the 1997 IEEE Symposium on Information Visualization (InfoVis '97)*, page 100, Washington, DC, USA, 1997.

[JW02] G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *Proc. 2002 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, pages 538–543, Edmonton, Canada, July 2002.

[KB00] R. Kosla and H. Blockeel. Web mining research: A survey. *SIGKDD Explorations*, 1:1–15, 2000.

[Kei02] D.A. Keim. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):1–8, 2002.

[KK01] M. Kuramochi and G. Karypis. Frequent subgraph discovery. In *Proc. 2001 Int. Conf. Data Mining (ICDM'01)*, pages 313–320, San Jose, CA, Nov. 2001.

[Kle99] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, 1999.

[KM05] J.Z. Kolter and M.A. Maloof. Using additive expert ensembles to cope with concept drift. In *Proc. 2004 Int. Conf. Machine Learning (ICML'05)*, pages 449–456, 2005.

[LFY+06] C. Liu, L. Fei, X. Yan, J. Han, and S. P. Midkiff. Statistical debugging: A hypothesis testing-based approach. *IEEE Trans. Software Engineering*, 32:831–848, 2006.

[LH07] X. Li and J. Han. Mining approximate top-k subspace anomalies in multi-dimensional time-series data. In *Proc. 2007 Int. Conf. Very Large Data Bases (VLDB'07)*, Vienna, Austria, Sept. 2007.

[LHKG07] X. Li, J. Han, S. Kim, and H. Gonzalez. Roam: Rule- and motif-based anomaly detection in massive moving object data sets. In *Proc. 2007 SIAM Int. Conf. Data Mining (SDM'07)*, Minneapolis, MN, April 2007.

[LHL08] J.-G. Lee, J. Han, and X. Li. Trajectory outlier detection: A partition-and-detect framework. In *Proc. 2008 Int. Conf. Data Engineering (ICDE'08)*, Cancun, Mexico, April 2008.

[LHM98] B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In *Proc. 1998 Int. Conf. Knowledge Discovery and Data Mining (KDD'98)*, pages 80–86, New York, NY, Aug. 1998.

[LHW07] J.-G. Lee, J. Han, and K. Whang. Clustering trajectory data. In *Proc. 2007 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'07)*, Beijing, China, June 2007.

[LHY+08] Xiaolei Li, Jiawei Han, Zhijun Yin, Jae-Gil Lee, and Yizhou Sun. Sampling cube: A framework for statistical olap over sampling data. In *Proc. 2008 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'08)*, page to appear, Vancouver, Canada, June 2008.

[Liu06] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, 2006.

[MH01] H. Miller and J. Han. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, 2001.

[MM02] G. Manku and R. Motwani. Approximate frequency counts over data streams. In *Proc. 2002 Int. Conf. Very Large Data Bases (VLDB'02)*, pages 346–357, Hong Kong, China, Aug. 2002.

[Mut03] S. Muthukrishnan. Data streams: algorithms and applications. In *Proc. 2003 Annual ACM-SIAM Symp. Discrete Algorithms (SODA'03)*, pages 413–413, Baltimore, MD, Jan. 2003.

[MXC+07] Q. Mei, D. Xin, H. Cheng, J. Han, and C. Zhai. Semantic annotation of frequent patterns. *ACM Trans. Knowledge Disocvery from Data (TKDD)*, 15:321–348, 2007.

[MZ05] Qiaozhu Mei and ChengXiang Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 198–207, 2005.

[OHS05] Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. Prediction and ranking algorithms for event-based network data. *SIGKDD Explor. Newsl.*, 7(2):23–30, 2005.

[PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Technical Report*, Computer Science Dept, Stanford University, 1998.

[PMM+02] Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart J. Russell, and Ilya Shpitser. Identity uncertainty and citation matching. In *Advances in Neural Information Processing Systems 15 (NIPS'02)*, pages 1401–1408, Vancouver, Canada, Dec. 2002.

[QD07] X. Qi and B. D. Davison. Web page classification: Features and algorithms. In *Technical Report LU-CSE-07-010*, Computer Science and Engineering, Lehigh University, 2007.

[RC07] Raghu Ramakrishnan and Bee-Chung Chen. Exploratory mining in cube space. *Data Min. Knowl. Discov.*, 15(1):29–54, 2007.

[RD02] Mathew Richardson and Pedro Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *Advances in Neural Information Processing Systems 14*, pages 1441–1448. MIT Press, 2002.

[RHS01] J. F. Roddick, K. Hornsby, and M. Spiliopoulou. An updated bibliography of temporal, spatial, and spatio-temporal data mining research. In *Lecture Notes in Computer Science 2007*, pages 147–163, Springer, 2001.

[SC03] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003.

[SQCF05] Jimeng Sun, Huiming Qu, Deepayan Chakrabarti, and Christos Faloutsos. Relevance search and anomaly detection in bipartite graphs. *SIGKDD Explor. Newsl.*, 7(2):48–55, 2005.

[SS03] Daniela Stan and Ishwar K. Sethi. eid: a system for exploration of image databases. *Information Processing and Management,*, 39:335–361, 2003.

[SW05] Cees Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools Appl.*, 25(1), 2005.

[SZ04] D. Shasha and Y. Zhu. *High Performance Discovery In Time Series: Techniques and Case Studies*. Springer, 2004.

[SZHV04] Shashi Shekhar, Pusheng Zhang, Yan Huang, and Ranga Raju Vatsavai. Trends in spatial data mining. In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha, editors, *Data Mining: Next Generation Challenges and Future Directions*, pages 357–380. AAAI/MIT Press, 2004.

[TcZ+03] Nesime Tatbul, Uğur Çetintemel, Stan Zdonik, Mitch Cherniack, and Michael Stonebraker. Load shedding in a data stream manager. In *vldb'2003: Proceedings of the 29th international conference on Very large data bases*, pages 309–320. VLDB Endowment, 2003.

[Tuf01] E. R. Tufte. *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press, 2001.

[WFYH03] H. Wang, W. Fan, P. S. Yu, and J. Han. Mining concept-drifting data streams using ensemble classifiers. In *Proc. 2003 ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD'03)*, pages 226–235, Washington, DC, Aug. 2003.

[WIZD04] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Springer, 2004.

[WK05] J. Wang and G. Karypis. HARMONY: Efficiently mining the best rules for classification. In *Proc. 2005 SIAM Conf. Data Mining (SDM'05)*, pages 205–216, Newport Beach, CA, April 2005.

[WLX+07] T. Wu, X. Li, D. Xin, J. Han, J. Lee, and R. Redder. Datascope: Viewing database contents in google maps' way. In *Proc. 2007 Int. Conf. Very Large Data Bases (VLDB'07)*, Vienna, Austria, Sept. 2007.

[WM03] T. Washio and H. Motoda. State of the art of graph-based data mining. *SIGKDD Explorations*, 5:59–68, 2003.

[WWYY02] H. Wang, W. Wang, J. Yang, and P. S. Yu. Clustering by pattern similarity in large data sets. In *Proc. 2002 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'02)*, pages 418–427, Madison, WI, June 2002.

[XCYH06] D. Xin, H. Cheng, X. Yan, and J. Han. Extracting redundancy-aware top-k patterns. In *Proc. 2006 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'06)*, pages 444–453, Philadelphia, PA, Aug. 2006.

[YH07] X. Yan and J. Han. Discovery of frequent substructures. In *D. Cook and L. Holder (eds.), Mining Graph Data*, pages 99–115, John Wiley Sons, 2007.

[YHY05] X. Yin, J. Han, and P. S. Yu. Cross-relational clustering with user's guidance. In *Proc. 2005 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'05)*, pages 344–353, Chicago, IL, Aug. 2005.

[YHY06] X. Yin, J. Han, and P. S. Yu. Linkclus: Efficient clustering via heterogeneous semantic links. In *Proc. 2006 Int. Conf. on Very Large Data Bases (VLDB'06)*, Seoul, Korea, Sept. 2006.

[YHY07a] X. Yin, J. Han, and P. S. Yu. Object distinction: Distinguishing objects with identical names by link analysis. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.

[YHY07b] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. In *Proc. 2007 ACM SIGKDD Int. Conf. Knowledge Discovery in Databases (KDD'07)*, San Jose, CA, Aug. 2007.

[YHYY06] X. Yin, J. Han, J. Yang, and P. S. Yu. Efficient classification across multiple database relations: A crossmine approach. *IEEE Trans. Knowledge and Data Engineering*, 18:770–783, 2006.

[ZYH+07] F. Zhu, X. Yan, J. Han, P. S. Yu, and H. Cheng. Mining colossal frequent patterns by core pattern fusion. In *Proc. 2007 Int. Conf. Data Engineering (ICDE'07)*, Istanbul, Turkey, April 2007.