

Automated, scalable systems would reveal and help exploit the deeper meanings in scientific data, especially in biomedical engineering, telecommunications, geospatial exploration, and climate and Earth ecosystem modeling.

EMERGING SCIENTIFIC APPLICATIONS IN DATA MINING



Recent progress in scientific and engineering applications has accumulated huge volumes of high-dimensional data, stream data, unstructured and semi-structured data, and spatial and temporal data. Highly scalable and sophisticated data mining tools for such applications represent one of the most active research frontiers in data mining. Here, we outline the related challenges in several emerging domains.

Biomedical Engineering

Biology is in the midst of a revolution, with an unprecedented flood of data forcing biologists to rethink their approach to scientific discovery. First, large-scale data-collection techniques have emerged for a number of data sources limited by throughput, or the amount of available data. Examples of the data glut include: systematic genome DNA sequencing of organisms; high-throughput determination of small molecule structures, as well as large macromolecular structures (such as proteins, RNA, and DNA); large-scale measurements of molecular interactions; and simultaneous measurement of the expression level of all genes (thousands to tens of thousands) in a population of cells. Second, the availability of this data requires biologists to create systems for organizing, storing, and disseminating it, thus creating a need for standard terminologies and the development of stan-

dards for interchange and annotation. Third, because of the apparent opportunities for automated learning from the data sets, a market for robust machine learning and data mining algorithms has emerged to take advantage of previous knowledge, without being overly biased in the search for new knowledge. As a result, biology has changed from a field dominated by an attitude of “formulate hypothesis, conduct experiment, evaluate results” to more of a big-science attitude of “collect and store data, mine for new hypotheses, confirm with data or supplemental experiment.” The long-term significance of the new data of molecular biology is that it can be combined with clinical medical data to achieve a higher-resolution understanding of the causes for and treatment of disease. A major challenge for data mining in biomedicine is therefore the organization of molecular data, cellular data, and clinical data in ways allowing them to be integrated for the sake of knowledge extraction.

A major additional source of information is the published medical literature, increasingly available online in full-text form or as useful (but unstructured) summaries of the main data and biomedical hypotheses.

Telecommunications

Data mining flourishes in telecommunications due to the availability of vast quantities of high-quality data. A significant stream of it consists of call records collected at network switches used primarily for

TERRY MUIRA

BY JIAWEI HAN, RUSS B. ALTMAN, VIPIN KUMAR, HEIKKI MANNILA, AND DARYL PREGIBON

billing; it enables data mining applications in toll-fraud detection [1] and consumer marketing [2].

Perhaps the best-known marketing application of data mining, albeit via unconfirmed anecdote, concerns MCI's "Friends & Family" promotion launched in the domestic U.S. market in 1991. As the anecdote goes, market researchers observed relatively small sub-graphs in this long-distance phone company's large call-graph of network activity, revealing the promising strategy of adding entire calling circles to the company's subscriber base, rather than the traditional and costly approach of seeking individual customers one at a time. Indeed, MCI increased its domestic U.S. market share in the succeeding years by exploiting the "viral" capabilities of calling circles; one infected member causes others to become infected. Interestingly, the plan was abandoned some years later (not available since 1997), possibly because the virus had run its course but more likely due to other competitive forces.

In toll-fraud detection, data mining has been instrumental in completely changing the landscape for how anomalous behaviors are detected. Nearly all fraud detection systems in the telecommunications industry 10 years ago were based on global threshold models; they can be expressed as rule sets of the form "If a customer makes more than X calls per hour to country Y , then apply treatment Z ." The placeholders X , Y , and Z are parameters of these rule sets applied to all customers.

Given the range of telecommunication customers, blanket application of these rules produces many false positives. Data mining methods for customized monitoring of land and mobile phone lines were subsequently developed by leading service providers, including AT&T, MCI, and Verizon, whereby each customer's historic calling patterns are used as a baseline against which all new calls are compared. So, for customers routinely calling country Y more than X times a day, such alerts would be suppressed, but if they ventured to call a different country Y' , an alert might be generated.

Methods of this type were presumably in place for the credit card industry a few years before emerging in telecom. But the size of the transaction streams are far greater in telecom, necessitating new approaches to the problem.

It is expected that algorithms based on call-graph analysis and customized monitoring will become more prevalent in both toll-fraud detection and marketing of telecommunications services. The emphasis on so-called "relational data" is an emerging area for data mining research, and telecom provides relational data of unprecedented size and scope.

These applications are enabled by data from the billing stream. As the industry transforms itself from a circuit-switched to a packet-switched paradigm, the data mining community could well experience a dearth of data, since billing is likely to be increasingly insensitive to usage. Moreover, the number of records that could potentially be recorded in a packet-switched network (such as packet headers) is orders of magnitude greater than today's circuit-switched networks. Thus, unless a compelling business need is identified, the cost of collecting, transmitting, parsing, and storing this data will be too great for the industry to willingly accept. A dearth of data could well spell the end to future significant data mining innovations in telecommunications.

However, this view might yet be altered by the following scenarios:

New network architectures. New-generation network infrastructure will have to adapt to changes in demand yet be more reliable and secure; for example, capacity in mobile networks will have to be assigned dynamically, necessitating development of new data mining techniques for understanding and predicting network load. Similarly, network-intrusion detection [3] will continue to be important to data mining, helping ensure that artificially induced traffic cannot cripple a network.

Mobility and microbilling. In Europe and Japan, merchants use mobile handsets for selling (and billing) a variety of consumer goods, including vending machine purchases and parking fees. Since these consumer activities correspond to "billable events," data will certainly be collected and maintained for such services.

Mobile services. Ease of use is crucial for enticing customers to adopt new mobile services. Data mining will probably play a major role in the design of adaptive solutions enabling users to obtain useful information with relatively few keystrokes.

Homeland security. Federal regulations require U.S. telecommunications companies to maintain call records for two years. With the recent emphasis on homeland security, along with the role telecom data can play identifying and tracking terrorist cells and activities, data will continue to be collected and maintained, even though the records may not be used for billing.

Geospatial Data

The scope, coverage and volume of digital geographic data sets have grown rapidly in recent years

due to the progress in data collection and data processing technologies. These data sets include digital data of all sorts, created, processed, and disseminated by government- and private-sector agencies on land use and socioeconomic infrastructure; vast amounts of georeferenced digital imagery and video data acquired through high-resolution remote sensing systems and other monitoring devices; geographic and spatiotemporal data collected by global positioning systems, as well as other position-aware devices, including cellular phones, in-vehicle navigation systems, and wireless Internet clients; and digital geographic data repositories on the Web. Moreover, information infrastructure initiatives, including the U.S. National Spatial Data Infrastructure, facilitate data sharing and interoperability, making enormous amounts of space-related data sharable and analyzable worldwide.

The increasing volume and diversity of digital geographic data easily overwhelm traditional spatial analysis techniques that handle only limited and homogeneous data sets with high-computational burden. To discover new and unexpected patterns, trends, and relationships embedded within large and diverse geographic data sets, several recent studies of geospatial data mining [4] have developed a number of sophisticated and scalable spatial clustering algorithms, outlier analysis techniques, spatial classification and association analysis methods, and spatial data-cleaning and integration tools.

Nevertheless, considering the challenges posed by the already enormous and increasing amount of spatial data, geospatial data mining is in its infancy. Lots of research needs to be done, especially concerning the following pressing issues:

Developing and supporting geographic data warehouses. Although data warehouses are central to the knowledge discovery process, no true geospatial data warehouse exists today. Creating one requires solutions to problems in geographic and temporal data compatibility, including reconciling semantics, referencing systems, geometry, accuracy, and precision. Creating a warehouse might also need to solve the problems of efficient computation of sophisticated spatial aggregations, as well as how to handle spatial-related data streams. However, spatial data warehouses are likely to eventually play an essential role in geospatial information exchanges and data mining, so it is critical that we develop and support such an infrastructure today.

Exploring and mining richer geographic data types. Geographic data sets are moving beyond the well-structured vector and raster formats to include semi-structured and unstructured data, especially

georeferenced stream data and multimedia data. Techniques have to be developed to handle spatiotemporal data, robust geographic concept hierarchies and granularities, and sophisticated geographic relationships, including non-Euclidean distances, direction, connectivity, attributed geographic space (such as terrain), and constrained interaction structures (such as networks).

Reaching a broader user community. Geospatial data mining needs to go beyond researchers to also deliver its benefits to general users. This requires high-level user interfaces and visualization tools that aid diverse users in geospatial data mining. Moreover, these interfaces and tools have to be integrated with existing geographical information systems and database systems to guide users searching for geographic knowledge, interpreting and visualizing discovered knowledge, and using the discovered geographic knowledge in their decision making.

Climate Data and the Earth's Ecosystems

The large amount of climate data acquired through NASA's Earth-observation satellites, terrestrial observations, and ecosystem models offers an unprecedented opportunity for predicting and preventing future ecological problems by managing the ecology and health of the planet. Such data consists of a sequence of global snapshots of the Earth, typically available at monthly intervals, including various atmospheric, land, and ocean variables (such as sea surface temperature, precipitation, and net primary production, or the net photosynthetic accumulation of carbon by plants). Due to the nature and scale of this data, data mining techniques can play a major role in the automatic extraction and analysis of interesting patterns, thus complementing existing statistical techniques.

Earth science data mining consists of two main components: the modeling of ecological data; and the design of efficient algorithms for finding spatiotemporal patterns. An important goal is the discovery of teleconnection patterns, or recurring and persistent climate patterns spanning vast geographical regions. They manifest themselves as spatiotemporal relationships among ecological variables observed at various locations on the Earth and are critical for understanding how the ecosystem's various elements interact with one another. Clustering techniques, which divide data into meaningful or useful groups, help automate the discovery of teleconnections [5, 6]. Specifically, clustering identifies regions of the Earth whose constituent points have similar short- and long-term climate characteristics. By analyzing correlations

among climate variables across these regions, it is possible to rediscover existing patterns (such as the El Niño periodic ocean-atmosphere disruption in the tropical Pacific Ocean), as well as new, previously unrecognized teleconnections. An alternative approach is to convert the time series into sequences of events, then apply existing association-rule techniques to discover interesting patterns in the sequences.

The difficulty of mining Earth science data is illustrated by the following examples of issues arising during the various stages of data mining analysis:

Preprocessing. It is often beneficial to aggregate data into a smaller number of points, easing computational requirements and (typically) reducing the amount of noise. However, it can be difficult for researchers to choose the proper level of aggregation, since too much limits the patterns that can be detected, while too little results in noisy data in which only the strongest patterns can be discovered. Event definition is another necessary but ill-defined task. In the spatial domain, the problem is too many events, and in the temporal domain, events are rare; for example, El Niño events occur only every four to seven years. Yet another concern is integrating data from heterogeneous sources (such as data covering different time periods). Earlier data may come from manual, Earth-based observations, while later data may originate from satellites.

Similarity of time series. The “proper” measure of similarity between time series is fraught with challenges. Linear correlation works well with standard clustering algorithms and lends itself to statistical tests. Nevertheless, alternate measures of time series similarity would be beneficial if they allowed the detection of patterns that could not be detected via linear correlation, and might, for example, be based on either dynamic time warping or cepstral coefficients representing the frequency spectrum of a time series. An “ideal” measure of similarity would account for time lag and the fact that only extreme events are usually correlated.

Identifying interesting patterns. Once patterns are discovered, it is difficult to distinguish the spurious ones from the significant ones. For example, given 40,000 time series recording the sea surface temperature at various points on the ocean’s surface and 60,000 time series representing precipitation on land, some of these series might, just by chance, have strong correlations. While a number of statistical approaches estimate significance levels, it is not possible to apply such approaches directly due to spatial and temporal autocorrelation. When genuine patterns are identified, domain-specific knowledge is inevitably still

needed to identify patterns of interest to Earth scientists.

Conclusion

These emerging applications involve great data-management challenges that also represent new opportunities for data mining research. Methods for mining biomedical, telecommunication, geospatial, and climate data are under active development. However, in light of the tremendous amount of fast-growing and sophisticated types of data and comprehensive data analysis tasks, data mining technology may be only in its infancy, as the technology is still far from adequate for handling the large-scale and complex emerging application problems. Research is needed to develop highly automated, scalable, integrated, reliable data mining systems and tools. Moreover, it is important to promote information exchange among users, data analysts, system developers, and data mining researchers to facilitate the advances available from data mining research, application development, and technology transfer. ■

REFERENCES

1. Fawcett, T. and Provost, F. Adaptive fraud detection. *Data Min. Knowl. Disc.* 1 (1997), 291–316, 1997.
2. Lambert, D. and Pinheiro, J. Mining a stream of transactions for customer patterns. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2001)* (San Francisco, Aug. 26–29). ACM Press, New York, 2001, 305–310.
3. Lee, W., Stolfo, S., and Mok, K. Adaptive intrusion detection. *Artif. Intell. Rev.* 14 (2000), 533–567.
4. Miller, H. and Han, J. *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis, London, U.K., 2001.
5. Kumar, V., Steinbach, M., Tan, P., Klooster, S., Potter, C., and Torregrossa, A. Mining scientific data: Discovery of patterns in the global climate system. In *Proceedings of the Joint Statistical Meetings* (Athens, GA, Aug. 5–9). American Statistical Association, Alexandria, VA, 2001.
6. Steinbach, M., Tan, P., Kumar, V., Klooster, S., and Potter, C. Data mining for the discovery of ocean climate indices. In *Proceedings of the 5th Workshop on Scientific Data Mining (SDM 2002)* (Arlington, VA, Apr. 13). Society of Industrial and Applied Mathematics, 2002, 7–16.

JIawei HAN (hanj@cs.uiuc.edu) is a professor in the Department of Computer Science in the University of Illinois at Urbana-Champaign.

Russ B. ALTMAN (altman@stanford.edu) is an associate professor of genetics and medicine (and computer science, by courtesy) in the Department of Genetics, Stanford Medical Informatics, Stanford University, Stanford, CA.

VIPIN KUMAR (kumar@cs.umn.edu) is Director of the Army High Performance Computing Research Center and a professor in the Computer Science Department at the University of Minnesota, Minneapolis.

HEIKKI MANNILA (Heikki.Mannila@cs.helsinki.fi) is Research Director of the HIIT Basic Research Unit in the Department of Computer Science at the University of Helsinki, Finland.

DARYL PREGIBON (daryl@research.att.com) is a member of AT&T Labs, Research, Florham Park, NJ.