Introduction

Jian Qin and M. Jay Norton

In the past few years, a number of research journals in library and information science have published review articles or special issues on knowledge discovery and data mining (Raghavan et al., 1998; Trybula, 1997; Vickery, 1997). These publications have primarily discussed background, scope and terminology, methods and techniques, and tools related to the topic from orientations other than library and information science. Research publications in library and information science have been implicitly related to knowledge discovery in databases (KDD) in terms of methods and techniques, though many of them did not use the terminology "knowledge discovery in databases" explicitly. This issue is devoted to aspects of KDD that are relevant or reflective of the field of library and information science.

Knowledge discovery in databases uses a variety of methods to evaluate data for relevant relationships that could yield new knowledge. According to Fayyad et al. (1996): "KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process" (p. 39). Data mining essentially focuses on identifying patterns previously not recognized and is considered only one component of the discovery process. KDD encompasses a growing collection of techniques, from a variety of disciplines, for investigating data to extract knowledge. The methods employ a broad combination and application of human expertise and information technology. "KDD comprises

Jian Qin, School of Information Studies, Syracuse University, 4-206 Center for Science and Technology, Syracuse, NY 13244
M. Jay Norton, School of Library and Information Science University of Southern Mississippi, Hattiesburg, MS 39406-5746
LIBRARY TRENDS, Vol. 48, No. 1, Summer 1999, pp. 1-8

many steps, which involve data preparation, search for patterns, knowledge evaluation, and refinement, all repeated in multiple iterations" (Fayyad et al., 1996, p. 41). KDD investigates databases to identify patterns of association, clusters, and rules but it requires significant rigor—not all patterns are real or meaningful. The presence of patterns may be meaningless and statistically insignificant. The successful use of data mining in KDD involves "data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of results of mining" (Fayyad et al., 1996, p. 39).

On a fundamental level, library and information services have been involved in component processes similar to the current definition of KDD. Practitioners and researchers in library and information science have expended significant resources—intellectual and physical—on investigating and developing methods to identify and exploit patterns within information entities. These methods are used to generate classification schemes and organization systems for information retrieval and to address often poorly expressed information needs of users. In seeking ways to provide better access to information, the field has attempted to determine characteristics of relevance in query construction and investigated methods for improving document retrieval. KDD studies in library and information science that Fayyad (1996) identifies relate to and drive KDD and include statistics, areas of artificial intelligence, pattern recognition, visualization, intelligent agents for distributed and multimedia environments, machine learning, databases, management information systems, knowledge acquisition, information retrieval, and digital libraries (p. 23). In some fields, KDD is interpreted as applying whatever computer rigor and capability is available for extracting information from databases of all constructions, while in others it may have fewer technological implementations but the same desired outcome—i.e., the discovery of useful information (Fayyad, 1996). Practitioners of library and information science may see themselves more as intermediaries, or part of the process, though researchers in the field may see themselves as discoverers. KDD is, and will continue to be, a complex, multidisciplinary, interdisciplinary arena requiring both practitioners and researchers. As the field continues to develop, it will be interesting to compare the disjointed records of some of the disciplines to determine if the same issues arise—i.e., standardization of database construction, development of algorithm rules related to specific topic collections, and questions of subject expert classification versus external classification systems.

The thirteen articles included in this issue characterize a combination of the knowledge discovery in data process components; the emerging information technology; and the established information methods such as classification, citation analysis, and indexing and abstracting. Norton's article begins the issue by giving an overview of what KDD is and what problems researchers face in KDD applications. She reviews the relationship between databases and knowledge discovery and the factors affecting the database quality that in turn impact the reliability and validity of KDD results. The article emphasizes that KDD is not at all a finished product, nor is it a panacea for all the research interests or ills of the database universe. In the face of many challenges in KDD, human involvement plays a vital role in the process.

Kwasnik discusses the relationship between knowledge representation (as manifested in classifications) and the processes of knowledge discovery and creation. While classifications categorize and interrelate domains and branches in the knowledge system, the classification process has potential to enable or constrain knowing something or discovering new knowledge about something. To demonstrate this, Kwasnik first describes the structures of a classification, including hierarchies, trees, paradigms, and faceted analysis with the goal of identifying how these structures serve as knowledge representations and in what ways they can be used for knowledge discovery and creation. When one considers that classification is built on known information, then KDD and classification takes on a new construction. Since a large part of KDD attempts to identify information that has previously been overlooked or unavailable, KDD will in itself affect classification. Basic constructs will remain the same but the underlying knowledge foundations that we apply to classification of an information entity will have to become more fluid in order to serve and be served by KDD. Kwasnik concludes that classification systems that are too rigid will not be applicable in the long term and may actually be detrimental to future knowledge discovery.

Swanson and Smalheiser report in their article the recent development in their text linkage discovery tool, Arrowsmith, a software program that draws upon expert knowledge in discovering implicit links among documents that have accumulated from the research done by Swanson (for a list of publications, see Swanson and Smalheiser's article in this issue of *Library Trends*) for more than a decade. Swanson's theory is based on the analogy that, if an article reports an association between substance *A* and some physiological parameter *B* while another reports a relationship between *B* and disease *C*, and a link between *A* and *C* via *B* has not been published previously, then to bring together the separate articles on *A-B* and *B-C* may suggest a novel *A-C* relationship of scientific interest. Arrowsmith is designed to develop systematic methods for discovering the undiscovered implicit relationships within the biomedical literature. It filters the text, matches phrases and concepts, and identifies potentially complementary items as pairs, which the researcher then analyzes for possible relationships. The software enables the investigator to evaluate relatively large bodies of data from a variety of aspects in a knowledge discovery mission. Recurrent in this work is the role of the human investigator—Arrowsmith is a tool for discovery but is not the discoverer. Software such as Arrowsmith enlarges the scope of our view but does not replace the human analysis. The ability to scrutinize substantial databases to extract potentially revealing, and previously unnoted, information is a result of improving technology that portends tremendous benefits.

The discussion by Cory describes his experiment using Swanson's methodology to investigate, through document retrieval, whether three philosophers from different times in history were influenced by one another. Discovering the undiscovered text linkages among documents is less problematic in the biomedical literature than in the humanities because the technical terminology is usually explicit and precise while the humanities literature often abounds with synonyms. Will Swanson's methods be applicable to the humanities literature and yield the same type of links among the humanities documents? The literature inquiry about three philosophers from different historical times showed evidence of influences that the third philosopher received from the second and the second received from the first. The search was able to identify publications showing that the third philosopher was also influenced by the first. While Cory's method departed significantly from Swanson's work due to the idiosyncrasies of the humanities language, this experiment nevertheless linked logically-related citations that were bibliographically unlinked. His article also discusses the problems in discovering hidden knowledge in humanities databases because of the nature of humanities research and the language used in the humanities literature.

Small demonstrates in his article how citation links can be used to map scientific passages crossing disciplinary boundaries. Both Swanson and Cory's studies indicate the importance of analogy in discovering covert relationships among documents. While their methodology focuses on "recurring" terms or names that are shared by the documents found, Small maintains that citation links represent a more direct author-selected dependency than vocabulary sharing. This allows citation links to be used to establish frequency patterns of co-citation or bibliographic coupling, and thus they are more objective in studying the unity of science from a global perspective. In his study, Small generated a path by selecting economics as the starting field and astrophysics as the destination field. The citation links reveal that this path traverses the fields of economics, psychology, neuroscience, biomedicine, genetics, chemistry, earth science, geoscience, semiconductors, lasers, and physics. The co-citation passage from economics to astrophysics embraces interdisciplinary boundary spanning, such as psychiatry to neuroscience, neuroscience to immunology, and biology to biochemistry.

A similar analogy to Swanson's can be made in co-citation passage analysis that, if *A* is in the starting field, *C* in the destination field, and *B* the shared concept/method, then *A* is to *B* as *C* is to *B*. Small suggests that, in future retrieval systems, a user could pick two topics or documents and generate a path of documents or topics that connect them, which could be used for information discovery and hypothesis generation.

The discussion by Qin addresses the problem of preprocessing and cleansing textual data for discovering semantic patterns in keyword frequency distributions. Keywords that are used as indexing terms in bibliographic records are semi-structured data. One challenge in mining such semi-structured data is to transform these into the types and structures suitable for statistical calculations and modeling. As semantic pattern analysis needs accurate data to draw valid and reliable conclusions, all the idiosyncrasies existing in natural language, including suffixes, different spellings for the same word, and synonyms, need to be normalized. Qin proposes the use of brief text codes to normalize the keywords while maintaining their original meaning. Besides the methodological aspect of mining bibliographic data, the frequency distribution patterns in the keyword data set suggest the existence of a common intellectual base with a wide range of specialties and marginal areas in the subject area studied. In normalizing the frequency of keyword occurrences, Qin found that the degree of keyword scattering at a certain region—i.e., keyword density—can be measured by the ratio of the number of unique keywords to the number of ranks at which the unique keywords occurred. The resulting values show a difference oftentimes between the specialty and marginal keyword regions. The semantic pattern analysis of the keywords from bibliographical coupling shows a possibility that simple semantic processing of natural language (keywords extracted from citation titles in this case) may be programmed into information retrieval tools for providing "analyzed" search results to users.

In his article, He reviews the development, applications, and advances made in co-word analysis during the last two decades. Though still developing as a technique, co-word analysis has been used in a variety of situations. Conjunct with its use is the recognition of one of its shortcomings—i.e., the assignment of keywords and indexing terms by indexers or database producers rather than the authors of the material. However, improving technology may allow the application of co-word analysis to full text to determine the appropriate keywords and indexing terms. It is through the application of such methods as co-word analysis that it is possible to identify problems in the construction of the databases and to consider the impact of indexers' choices on future retrieval and understanding of the semantic structures of a discipline. The creation of knowledge discovery methods also results in knowledge discovery as it highlights issues, concerns, and activities not previously scrutinized under other methods.

The articles mentioned above have concentrated on finding document content linkages and semantic patterns from the data available in bibliographic databases. As digital documents grow exponentially, needs for organizing and retrieving these documents also arise. How can the subject content of digital full-text documents be represented effectively for retrieval purposes? What characteristics exist in these digital documents? How can these characteristics be organized and implemented in information systems to assist people in knowledge discovery? The following contributions address these questions from three different perspectives.

Ahonen's article analyzes digital document collections by identifying descriptive or meaningful word sequences that may be used in a variety of knowledge discovery missions. In extracting frequent word sequences from full-text documents, Ahonen posits that there may be common measures of relevance that can be detected by examining characteristics of word sequences. Her discussion provides a detailed account of the methods involved and demonstrates the potential of word sequence evaluation for knowledge discovery. Patterns in word sequences may be produced, based on a combination of pre- and post-processing linked to the specific application and frequency relations defined by rule sets and weight systems. The patterns may suggest areas of further investigation, be used to preevaluate a document's relevance without examining the whole document, or provide context for one not familiar with the document collection. The subject expert might also discern new information from the sequence associations or patterns.

Chowdhury presents a selection of cases where template mining has been successfully applied for information extraction from digital documents. Additionally, he reports on template use in Web search engines conducting information retrieval rather than information extraction. The initial distinction is that information retrieval attempts to locate relevant documents from collections while information extraction attempts to pull relevant information from documents. Though these are degrees of retrieval to some, the difference can be significant. The templates designed to assist the Web user in searching are created by expert searchers who organize information into groups and topics that are used to create the template structure for the less experienced user to plug into. The template is used to locate documents. The templates he ultimately focuses on have the potential advantages of authors using the template system to implement a more controlled method for creating document surrogates and digital document description to better enable information extraction from the documents, not just the collection. Not proposing a single all-purpose metadata format at this time, he suggests further research and investigation into what would be the most appropriate format.

Desai et al. developed a virtual library indexing and discovery system named CINDI (Concordia INdexing and DIscovery System) that allows authors of digital documents to describe their document via completion of a semantic header and use of an expert registry subsystem. An appealing aspect of knowledge discovery in databases involves locating knowledge that might otherwise be overlooked. The Internet search engines often suffer from a lack of organization and consistency in the collection space. An extraordinary number of retrieved documents preclude appropriate evaluation and tend to result in missed opportunities rather than recovered data. Some of the more complex endeavors of KDD are seeking ways to access legacy data that are not organized consistently. Current developers of data warehouses are encouraging more standardization as future redress for the problem (Bontempo & Zagelow, 1998). The header contains the metadata used by the searching systems to determine the appropriateness of retrieving that resource. CINDI provides assistance comparable to the expert cataloger or indexer for the author, addressing the shortcomings of many current search engines via better metadata description. The outcome of the use of CINDI should be a significant improvement in the ability of searchers to locate materials relevant to their inquiry. This knowledge discovery approach begins with the initial document, which will produce improved results in the future. It will rely more on known relationships than unknown but should enhance retrieval of related documents.

Pinto and Lancaster offer a new view on abstracts and abstracting—i.e., that the quality of abstracts is extremely important in knowledge discovery tasks. Because of the dual roles of content descriptor and retrieval tool, abstracts must maintain the quality of accuracy, readability, cohesion/coherence, and brevity. However, the importance of these criteria is likely to vary depending on who will be reading the abstracts. For abstracts intended solely for search purposes, such criteria as readability and coherence/cohesion are not important, while other attributes are applicable in other ways. Pinto and Lancaster maintain that the increasing application of computers to text processing has not reduced the value of abstracts, and their value should not diminish as more critical or sophisticated operations, including those of knowledge discovery, are applied to the text.

In exploring knowledge from geospatial information systems (GIS), Yu demonstrates, through GeoMatch, a GIS-based prototype system for cartographic information retrieval, that coordinates data in MARC records can be processed to provide understandable and useful knowledge for users in selecting information relevant to their needs. GeoMatch is a graphic-based interface that mines the geographical data buried in MARC records and other geospatial sources and visualizes the new knowledge discovered in these data. Discovering knowledge in geospatial data is distinct from text information searching because it uses algorithms to convert the coordinates information into user-understandable and useful knowledge. The main contribution of GeoMatch is the quantitative analysis of overlapping relationships in the retrieval process. Not only can it help users to more precisely define their information need and adjust the searching strategy, but also it can be used to rank the result. The KDD applications of this type have constructive implications for information retrieval.

Finishing out this issue is distinguished Professor Emeritus Herbert S. White, former dean of Indiana University School of Library and Information Science. In "Librarians and Information Technology: Which is the Tail and which is the Dog?" he discusses the role of library professionals in relation to the applications of database technology. He argues that some information technology has positioned the librarian contrary to the supportive service role that has surrounded the profession.

REFERENCES

Bontempo, C., & Zagelow, G. (1998). The IBM Data Warehouse Architecture. *Communications of the ACM, 41*(9), 38-48.

Borgman, C. (1986). Why are online catalogs hard to use? Lessons learned from information-retrieval studies. *Journal of the American Society for Information Science, 37*(6), 387-400.

Fayyad, U. M. (1996). Data mining and knowledge discovery: Making sense out of data. *IEEE Expert, 11*(5), 20-25.

Fayyad, U. M., & Stolorz, P. (1997). Data mining and KDD: Promises and challenges. *Future Generation Computer Systems, 13*(2-3), 99-115.

Fayyad, U. M.; Piatetsky-Shapiro, G.; & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

Frawley, W. J.; Piatetsky-Shapiro, G.; & Matheus, C. J. (1991). Knowledge discovery in databases: An overview. In G. Piatetsky-Sharpiro & W. J. Frawley (Eds.), *Knowledge discovery in databases* (pp. 1-27). Cambridge, MA: AAAI Press.

Raghavan, V.V.; Deogun, J. S.; & Sever, H. (Eds.). (1998). Special topical issue: Knowledge discovery and data mining. *Journal of the American Society for Information Science, 49*(5).

Trybula, W. J. (1997). Data mining and knowledge discovery. *Annual Review of Information Science & Technology, 32,* 197-229.

Vickery, B. (1997). Knowledge discovery from databases: An introductory review. *Journal of Documentation, 53*(2), 107-122.