1. Introduction

Data mining techniques (DMT) have formed a branch of applied artificial intelligence (AI), since the 1960s. During the intervening decades, important innovations in computer systems have led to the introduction of new technologies (Ha, Bae, & Park, 2000), for web-based education. Data mining allows a search, for valuable information, in large volumes of data (Weiss & Indurkhya, 1998). The explosive growth in databases has created a need to develop technologies that use information and knowledge intelligently. Therefore, DMT has become an increasingly important research area (Fayyad, Djorgovski, & Weir, 1996).

Of the data mining techniques developed recently, several major kinds of data mining methods, including generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining, are herein reviewed. The techniques for mining knowledge from different kinds of databases, including relational, transactional, object oriented, spatial and active databases, as well as global information systems, are also examined. Potential data mining applications and some research issues are discussed.

As an element of DMT research, this paper surveys the development of DMT, through a literature review and the classification of articles, from 2000 to 2011. The various applications of DMT, during that period, are reviewed. This period is especially important, because the Internet was opened to general users, in 2000, and the newly widespread availability of information and communication technology has played an important role, not only in the field of DMT, but also in the development of methodologies for the collection of data from online databases.

The period of interest, for this literature survey, begins in January 2000. In August, 2011, a search was made of the keyword indices on the Elsevier SCOPUS, Springerlink, IEEE Xplore, EBSCO (electronic journal service) and Wiley InterScience online database, for article abstracts containing the phrase, “data mining technique”. For the period from 2000 to 2011, 14,972 articles were found. Topic filtering reduced this number to 216 articles, from 159 journals, which were related to the keyword, “Data mining application”. Using these 216 articles on DMT applications, this paper surveys and classifies DMT, using nine categories: Neural networks, Algorithm architecture, dynamic prediction-based, Analysis of systems architecture, Intelligence agent systems, Modeling, knowledge-based systems, System optimization and Information systems, together with their applications in different research and practical domains.

The remaining part of the paper is organized as follows. Sections 3–11 present the survey results for DMT methodologies and applications, based on the categories mentioned, above. Section 12 is a discussion of suggestions for the future development of DMT methodologies and their applications. Finally, Section 13 contains a brief conclusion.

2. Trend in data mining techniques

The data collected dates from 2000, until August 2011. The trend for the keywords; Data mining, Decision tree, Artificial
neural network, Clustering, Association rule, Artificial intelligence, Bioinformatics, Customer relationship, Fuzzy logic, and their applications, are shown in Table 1.

3. Neural networks and their applications

The term, neural network, is traditionally used to refer to a network, or circuit of biological neurons. Modern use of the term often refers to artificial neural networks, which are composed of artificial neurons, or nodes. As well as electrical signaling, other forms of signaling arise from neural transmitter diffusion, which have an effect on electrical signaling. As such, neural networks are extremely complex.


4. Algorithm architecture and its applications

Algorithm architecture is expressed as a finite list of well-defined instructions, to calculate a function. Algorithms are used for calculation, data processing and automated reasoning. Simply put, an algorithm is a step-by-step procedure for calculation. A partial formalization of the concept began with attempts to solve the Entscheidungs problem, posed by David Hilbert in 1928.

Some applications that are implemented by algorithms include gap statistic algorithms, chi-square automated interaction detection, models and algorithms, GRASP, OLAP, K-means, Clustering algorithms, decision forest algorithms, classification and regression trees, Euclidean distance, bagged clustering algorithms, fuzzy logic, association rules, C&RT, Apriori algorithms, CS, anomaly-based IDS, clustering, genetic algorithms, CRISP-DM models, thyroid stimulation and SVM. Algorithm architectures and their applications are listed in Table 3.

5. Dynamic prediction based approach and its applications

The dynamic prediction based approach is a mathematical model for stochastic dynamics; used in modeling molecules, but it also finds applications in the stock market, among other areas. The most important feature of Langevin dynamics is the presence of a Gaussian random noise. The principle of temporal locality was most important feature of Langevin dynamics; used in modeling molecules, but it also finds applications in the stock market, among other areas. The stock market, among other areas.

Some applications which use a dynamic prediction based approach include ophthalmic oncology, vehicle fault diagnosis, grid computing, dyadic wavelet, pre-fetching, fault restoration prediction models, fault prediction models, financial distress prediction models, Vlasov–Maxwell equations, chemical reactivity predictions, real time vehicle tracking, forecasting, anomaly detection, churn prediction, comparative genomics, clinical predictions and predictive models. Table 4 categorizes predictions into read-set predictions and write-set predictions.

6. Analysis of system architecture and its applications

The analysis of system architecture uses a conceptual model that defines the structure, behavior and other aspects of a system. Systems architecture makes use of elements of both software and hardware, which allows the design of composite systems. A good architecture may be viewed as a ‘partitioning scheme’, or algorithm, which completely partitions all of the system’s present and foreseeable requirements into a workable set of clearly bounded subsystems.

Some applications of the analysis of system architecture are semantic analysis, regression analysis, statistical analysis, discriminative analysis, association analysis, penalized discriminative analysis, process parameter analysis, cluster analysis, decision making, decision support systems, consumer behavior analysis, binary logistical regression analyses, M5 model trees, factor analysis, market basket analysis, collaborative filtering, data analysis, decision tree based models, principal component analysis, multi-feature selection, intrusion detection and hem dialysis. These system architectures and their applications are listed in Table 5.

7. Intelligence agent systems and their applications

In the field of artificial intelligence, an intelligent agent system (IAs) is an autonomous entity, which observes and acts upon an environment. Intelligent agents may also learn, or use knowledge to achieve their goals. They may be very simple, or very complex. A reflex machine, such as a thermostat is an intelligent agent, as is a human being, as is a community of human beings working together towards a goal. Russell and Norvig (2003) considered goal-directed behavior as the essence of intelligence and so borrowed the term, “rational agent”, from economics.

Some applications for intelligent agent systems include multi-agent systems, complex systems, computer interface design, multiluser database systems, intelligent analysis, manufacturing intelligence, intelligent tutoring systems, support vector machines, program diagnostics systems, supervisory and specialist systems, supervisory and specialist systems, computing intelligence, artificial intelligence and Mahalanobis Taguchi systems. These intelligent agent systems and their applications are detailed in Table 6.

Table 1

<table>
<thead>
<tr>
<th>Keyword</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>2011</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data mining</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>10</td>
<td>16</td>
<td>16</td>
<td>16</td>
<td>114</td>
</tr>
<tr>
<td>Decision tree</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>Artificial neural network</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>Clustering</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>Association rule</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Artificial intelligence</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Bioinformatics</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Customer relationship management</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Fuzzy logic</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>1</td>
<td>10</td>
<td>9</td>
<td>11</td>
<td>15</td>
<td>15</td>
<td>24</td>
<td>16</td>
<td>32</td>
<td>27</td>
<td>26</td>
<td>188</td>
</tr>
</tbody>
</table>
8. Modeling and its applications

Modeling, in software engineering, is the process of creating a data model by making descriptions of formal data models, using data modeling techniques. Modeling technology can provide quantitative methods for the analysis of data, to represent, or acquire expert knowledge, using inductive logic programming, or algorithms, so that AI, cognitive science and other research fields are afforded broader platforms for the development of DMT.

Applications of modeling include cost modeling, model-based diagnosis, forest fire proliferation modeling, model output statistics, intonation modeling, XML document modeling, Cox proportional hazard modeling, load damage exponents, polynomials, similar waveforms, simple additive weight, computer numerical control, meta learning and drug utilization. The applications of modeling techniques are listed in Table 7.

9. Knowledge-based systems and their applications

Knowledge-based systems are artificial intelligent tools that work in a narrow domain, to provide intelligent decisions, with justification. The most common definition of KBS is human-centered, since KBS have their roots in the field of artificial intelligence (AI). They represent attempts to understand and initiate human knowledge, in computer systems (Wiig, 1994).
Knowledge is acquired and represented, using various knowledge representation techniques, rules, frames and scripts. The basic advantages offered by such systems are the documentation of knowledge, intelligent decision support, self-learning, reasoning and explanation. Akerkar and Saja Priti Srinivas (2009) stated that knowledge-based systems are systems based on the methods and techniques of Artificial Intelligence.

Some applications of knowledge-based systems include learning techniques, auto control techniques, knowledge discovery in databases, knowledge spirals, communication technologies, knowledge measurement, knowledge extraction, knowledge acquisition, knowledge management, knowledge representation, digital libraries and information gain theory data mining. The applications of knowledge-based systems are listed in Table 8.

10. System optimization and its applications

Fermat and Lagrange used calculus-based formulas, for the identification of optima, while Newton and Gauss proposed iterative methods of approaching an optimum. Historically, the original term for optimization was “linear programming”, coined by George B. Dantzig, although much of the theory had been described by Leonid Kantorovich, in 1939.

Dantzig published the Simplex algorithm, in 1947, and developed the theory of duality, in the same year. System Optimization refers to the selection of a best element, from some set of available alternatives. In the simplest case, problems in which a real function is maximized, are solved by systematically choosing the values of real or integer variables, from within an allowed set.

Some applications of system optimization include electrical nerve stimulation, R-peaks detection, latent reference individual extraction methods, operation optimization values, vertical partitioning, logistical regression, analytical hierarchy processes, polynomial regression, biogeography based optimization, particle swarm optimization, finite element methods, discrete rough set methods, asymptotic methods and parallel computing. The applications of system optimization techniques are listed in Table 9.

11. Information systems and their applications

Information systems are the products of an academic discipline. They occupy a place between the business world and computer science bridging the business field and the well-defined computer science field that is evolving toward a new scientific area of study. An information system relies on the theoretical foundations of information and computing which allows researchers a unique opportunity to engage in academic studies of various business models and related algorithmic processes that are pertinent to computer science. In general, information systems focus upon the processing of information within organizations (Hoganson, 2001), especially within business enterprises. The products of the process can then be share with society.

Some applications of information systems include patient characteristics, catchment characteristics, mobile databases, self-organizing, feature maps, insurance claims databases, alternating current field measurement, fracture acidizing, latest time sub-series, destination choice, attribute relevance studies, fraudulent financial statements, sequence similarity, case-based reasoning, anthropometric data, regression spines, economic imbalances, medium-voltage customer faults, maintenance and engineering, bank lending, reinforcement learning, supervised learning, arousals, information visualization, customer retention, churn management, pattern discovery, customer relationship management and uniaxial compressive strength. The applications of information systems are listed in Table 10.

<table>
<thead>
<tr>
<th>Table 8</th>
<th>Knowledge-based systems and their applications.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge-based systems applications</td>
<td>Authors</td>
</tr>
<tr>
<td>Data analysis techniques</td>
<td>Górecki-Zimroz et al. (2005)</td>
</tr>
<tr>
<td>Learning techniques</td>
<td>Chun and Park (2006)</td>
</tr>
<tr>
<td>Auto control techniques</td>
<td>Li et al. (2006)</td>
</tr>
<tr>
<td>Knowledge discovery in databases</td>
<td>Wason et al. (2006)</td>
</tr>
<tr>
<td>Knowledge spiral</td>
<td>Wason et al. (2006)</td>
</tr>
<tr>
<td>Communication technologies</td>
<td>Ko and Osei-Bryson (2006)</td>
</tr>
<tr>
<td>Knowledge measurement</td>
<td>Shapiro and Youtie (2006)</td>
</tr>
<tr>
<td>Knowledge extraction</td>
<td>Sugihara (2006)</td>
</tr>
<tr>
<td>Knowledge acquisition</td>
<td>Gajzler (2010)</td>
</tr>
<tr>
<td>Knowledge management</td>
<td>Fesharaki et al. (2011)</td>
</tr>
<tr>
<td>Knowledge representation</td>
<td>Hoonakker et al. (2011)</td>
</tr>
<tr>
<td>Digital libraries</td>
<td>Segura et al. (2011)</td>
</tr>
<tr>
<td>Information gain theory data mining</td>
<td>Sudheep et al. (2011)</td>
</tr>
</tbody>
</table>

12. Discussions, limitations, and suggestions

12.1. Discussions

A top-level analysis of data mining technologies must focus on data retention. In early attempts at data mining, the data set was maintained, for future pattern matching. This literature review shows that DMT application and development has diversified, in line with the various authors’ backgrounds, expertise, and areas of interest. For this reason, some authors are associated with literature concerning more than one methodology, or application.

It is also true that some techniques have common concepts and types of methodology. For example, dynamic prediction based methods, knowledge-based systems and neural network applications. A few authors are associated with literature concerning different methodologies and applications. This indicates that the development of methodologies is also diverse, in accordance with each author’s research interests and areas of interest. This seems to indicate that the development of DMT more expertise oriented.

Furthermore, some applications afford a greater opportunity for the use of different methodologies. For example, neural classification, Bayesian confidence propagation neural networks, gene regulatory networks, fuzzy recurrent neural networks, C4.5, Apriori algorithms, C5, anomaly-based IDS, clustering, genetic algorithms, CRISP-DM models, thyroid stimulation, Vaslov–Maxwell equations, chemical reactivity predictions, real time vehicle tracking, forecast, anomaly detection, churn prediction, knowledge measurement, knowledge extraction, knowledge acquisition, knowledge manage-
ment and knowledge representation are all topics with different methodologies, which nevertheless implement DMT, for problems that are common to all. These applications represent a major part of DMT development, but many methodologies are used to solve the problems that are specific to them. This would seem to imply that future developments of DMT will be more problem-centered.

In this paper, the articles discussed were sourced from different discipline areas, including computer science, engineering, medicine, mathematics, earth and planetary sciences, biochemistry, genetics and molecular biology, business, management and accounting, social sciences, decision sciences, multidisciplinary, environmental science, energy, agricultural and biological sciences, nursing, materials science, pharmacology, toxicology and pharmaceutics, chemistry, health professions, physics and astronomy, economics, econometrics and finance, psychology, neuroscience, chemical engineering and veterinary, which were all retrieved from the Elsevier SCOPUS, Springerlink, IEEE Xplore, EBSCO (electronic journal service) and Wiley InterScience online database.

Industrial applications of data mining techniques have increased, between 2000 and 2011. Fig. 1 shows the important DMT trends for association rules, genetic algorithms, clustering, artificial neural networks, Apriori algorithms, support vector machines, feature selection, customer relationship management, classification, neural networks and decision trees.

It cannot be concluded that DMT methodologies and applications are not developed in other science fields. However, studies that mention more applications of DMT in different research fields must be published, in order to broaden the scope of DMT, in the academic and practical fields.

### 12.2. Limitations

This study has some limitations. Firstly, a widespread literature review of DMT and its applications presents a difficult task, because of the extensive background knowledge that is required, when collecting, studying and classifying these articles. Although acknowledging a limited background knowledge, this paper makes a brief review of literature concerned with DMT, from 2000 to 2011 in order to determine how DMT and their applications have developed, in this period. Indeed, the categorization of methodologies and their applications is based on the keyword index and article abstracts, collected for this research. Some other articles may have used similar DMT methodologies in their applications, but may not have a DMT index, so this paper is unaware of these reference sources. Therefore, the first limitation of this article is the author’s limited knowledge of this subject. Secondly, although 216 articles from 159 academic journals (five online databases) are cited in this paper, other academic journals are listed in the science citation index (SCI) engineering index (EI) and the social science citation index (SSCI), as well as other academic journals/magazines, practical articles and reports that are not included in this survey. These would have provided more complete information about the developments in DMT and their applications. Lastly, non-English publications were excluded from this study. It is believed that research regarding the application of data mining techniques has also been discussed and published in other languages.

### 12.3. Suggestions

1. Other social science methodologies. This article’s definition of DMT is not complete, because other methodologies, such as social science methodologies, were not included in the survey. However, qualitative questionnaires and statistical methods are a type of research technology that is often used in social studies. For example, cognitive science, psychology and human behavior are used to implement different methods for investigating specific human problems, so other social science methodologies may include DMT in future studies.

2. Integration of methodologies. DMT is an interdisciplinary research topic, so future development of DMT must be integrated with different methodologies. This integration of methodologies and cross-disciplinary research may offer new insights into the problems associated with DMT.

3. Change is a source of future ES development. Change, due to social and technical reasons, can either enable, or inhibit ES methodologies and the development of applications. It can be seen that inertia, stemming from the use of routine problem solving procedures, stagnant knowledge sources and reliance on past experience, or knowledge may impede change, with respect to learning and innovation, for individuals and organizations.

Continued creation, sharing, learning and the acquisition of knowledge about different methodologies and applications also plays a key role in ES development.
13. Conclusions

This paper presents a review of literature concerned with DMT and its applications, from 2000 to 2011. It used a search of keyword indices and article titles. It is concluded that development of DMT is tending to become more expertise-oriented and that the development of DMT applications is more problem-centered. It is suggested that different social science methodologies, such as psychology, and cognitive science and human behavior might use DMT as an alternative methodology. Integration of qualitative, quantitative and scientific methods and the integration of studies of DMT methodologies will increase understanding of the subject. Finally, the ability to continually change and provide new understanding is the principal advantage of DMT methodologies, and will be at the core of DMT applications, in future.

References


