Data Mining for the Enterprise

Charly Kleissner, Ph.D.
Vice President, Engineering
Ariba Technologies
ckleissner@ariba.com

Abstract

The emergence of comprehensive data warehouses which integrate operational data with customer, supplier, and market data have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of that data. However, there has been a growing gap between more powerful data warehousing systems and the users' ability to effectively analyze and act on the information they contain. Data mining tools and services are providing the leap necessary to close this gap. Data mining offers automated discovery of previously unknown patterns as well as automated prediction of trends and behaviors; its technologies are complimentary to existing decision support tools and provide the business analyst and marketing professional with a new way of analyzing the business.

After a general introduction of the knowledge discovery lifecycle and the data mining lifecycle, this article examines the data mining issues and requirements within an enterprise. A comprehensive architectural overview proposes data mining integration solutions for data warehouses, application servers, thick clients, and thin clients. This article concludes with an analysis of current trends relevant to enterprise usage of data mining tools and methodologies.

1. Introduction

A number of business trends have made the usage of data mining tools and services mandatory for companies vying for business in today's competitive market place. The following section gives an overview of these business trends.

1.1. Business trends

- **Data explosion:**
  As companies are confronted with the challenge of handling an ever-increasing amount of data, it is becoming more difficult for business professionals to understand the desired *information* from this data. Data mining, which deals with the discovery of hidden knowledge, unexpected patterns, and new rules from large databases and data warehouses, promises to alleviate some of this difficulty.
- **Business reengineering and organizational decentralization:**
  Over the past few years, corporations have been reengineering their business processes and organizations. This has resulted in flatter and leaner organizations where knowledge workers have the authority and responsibility to implement and recommend business process optimizations and improvements.
- **Faster product cycles:**
  Most companies are challenged by the need for faster product and service development cycles in order to take advantage of newly emerging market opportunities. Data mining tools provide the means for proactively identifying new product opportunities and cross-selling products and services into existing customer accounts.
- **Globalization and enterprise topologies:**
  The globalization of the economy is in part supported and enabled by enterprise information system topologies where distributed computing is becoming the dominant computing paradigm. Desktop, client/server, transaction processing, object oriented infrastructures, and internet/intranet technologies are converging rapidly to create the environment required for enterprise-wide computing. Data mining technologies, methodologies, tools, and services need to take advantage of and add new components to this enterprise infrastructure.

Core components of data mining technology have been under development for decades in research areas such as statistics, artificial intelligence, and machine learning. Today, the maturity of these techniques, coupled with the evolution of massive data warehouses and sophisticated enterprise topologies make these technologies practical and ready for enterprise wide deployment in business

1060-3425/98 $10.00 (c) 1998 IEEE

applications. Integrated data warehousing solutions, data mining enhancements to desktop tools, and data mining services on application servers, provide a complete range of data mining capabilities aimed at empowering business professionals to more effectively make decisions and recommend business strategies.

1.2. Data mining definition

Data mining is a new decision support analysis process to find buried knowledge in corporate data and deliver understanding to business professionals. This definition contains some key phrases which deserve more elaboration:

- **Data mining is a *process*:**
  Data mining is not a one-time activity of a single business analyst, but rather a commitment of an organization to leverage its business data on an on-going basis and to continuously and iteratively improve its business practices based on a new level of understanding of its data.
- **Data mining is *complimentary* to decision support tools:**
  Today's decision support tools are *'assumption-driven'* tools in the sense that the business professionals use them to verify hypothesis about the data. Data mining tools are *'discovery-driven'* and complimentary to assumption-driven tools (see [2]). Data mining tools are hypothesis generators; they analyze corporate data automatically to discover new insight into business experiences.
- **Data mining finds *buried knowledge*:**
  Over the last few years data mining has sometimes been overhyped with respect to finding hidden *treasures* using these new algorithms and methodologies. Experience has shown that data mining is most effective in finding *buried knowledge* which provides additional insight into business practices.
- **Data mining delivers understanding to *business professionals*:**
  The new breed of data mining tools is geared towards business professionals, business analysts, and marketing professionals - not the statistician or artificial intelligence expert. Data mining tools do not necessarily replace statisticians, but they empower knowledge workers to more deeply understand business data without having to comprehend details of statistical analysis or intricacies of artificial intelligence.

1.3. Outline of article

The remainder of this article is structured as follows: chapters 3 and 4 provide a description of the knowledge discovery and data mining lifecycles. This provides the background for discussing data mining issues and requirements within an enterprise in chapter 5. Chapter 6 elaborates on various data mining integration strategies. Chapter 7 describes in more detail the various data mining architecture components, and their relationship to other heterogeneous distributed information services. A discussion of major data mining industry trends is presented in the final chapter.

2. Data mining solutions

The following two sections briefly describe data mining solutions which have been successfully developed and deployed over the last few years. Reference [17] surveys the growing number of industrial applications of data mining.

2.1. Vertical data mining solutions

- **Retailing:**
  In the retailing industry, data mining tools are used for market basket analysis (i.e., the process of determining which products a customer typically purchases at the same time) and targeted marketing campaigns.
- **Health care:**
  The health care sector uses data mining for patient behavior analysis and therapy analysis.
- **Banking and securities:**
  The banking industry is dependent on data mining tools for credit authorization (see [11]) and credit card fraud detection.
- **Insurance:**
  Insurance knowledge workers have successfully applied data mining techniques to claim analysis and fraud analysis.
- **Transportation:**
  Transportation industries have successfully used data mining for analyzing loading patterns.

2.2. Horizontal data mining solution

The prime example of a horizontal data mining solution is *customer lifecylce management*. Customer lifecycle management is not an industry specific solution, even though it may be customized for a given vertical market. Benefits of using data mining methodologies for customer management include increased market share, increased customer value, and decreased market loss.

The customer lifecycle management solution comprises applications for the following three phases of customer management:

- **Acquiring customers:**

Data mining has been successfully applied to finding new customers. A marketing professional first analyzes the existing customer base to understand which products are bought by which type of customer. This knowledge can then be used to conduct very targeted marketing campaigns for potential customers. Potential customers can be found by using the predictive capabilities of data mining against a set of purchased demographic data.

- **Increasing customer value:**
  Data mining has been effectively applied to cross-sell and launch new products to an existing customer base. A marketing professional first analyzes the existing customer base to understand which combination of products are bought by which type of customer. This knowledge can then be used to conduct very targeted marketing campaigns for existing customers who might benefit from additional products.
- **Retaining customers:**
  Data mining is necessary to retain high-valued customers for the long term and prevent them from signing up with the competition. A marketing professional first analyzes historical data of customers who have chosen a competitor's product. The predictive modeling capabilities can then be used to predict which existing customers are most likely to switch to a competitive product. Targeted marketing activities could help retain these customers thereby diminishing customer attrition.

3. Knowledge discovery lifecycle

The activity of data mining is part of a more global business process generally referred to as knowledge discovery process. The main phases of the knowledge discovery process are: data selection, data cleansing, data enrichment and coding, and data mining. We describe each of these stages in more detail in the following sections.

The knowledge discovery process is iterative in nature, e.g., based on some data mining results, a marketing professional might decide to invest more time and money to clean the data, or to enrich the original data with more information.

3.1. Data selection

Data selection deals with the assessment of which data will be used for data mining. In most cases, operational databases have not been designed to store historical data or to respond to queries, but rather to support all the applications for day-to-day transactions.

Therefore IT organizations have implemented data warehouses to support strategic decision support activities. For the purposes of this discussion, we consider the transfer of data from the operational databases to the data warehouse as part of the data selection phase.

3.2. Data cleansing

Some data cleansing activities might occur before the data mining process starts, either on the source data in the operational databases or as part of the data warehouse management and maintenance activities. Other data cleansing activities can only be initiated after pollution is detected as part of the data enrichment and coding phase or the initial data mining activities.

Typical cleansing activities include eliminating duplicate data entries and correcting inconsistent data.

3.3. Data enrichment and coding

Additional data which is not available in the operational databases or the data warehouses might be important for conducting meaningful data mining studies. This phase of the knowledge discovery lifecycle is referred to as the data enrichment phase.

Data in the data warehouse can benefit from a number of transformations which collectively are referred to as the data coding phase. This phase might include deleting some records (e.g., records which do not have enough meaningful information) and recoding or transforming other information (e.g., transforming addresses into regional codes, transforming purchase dates into numbers, transforming birth dates into age classes, etc.). Some of these transformations are best done in preparation for data mining, others as part of data mining.

At the end of the data enrichment and coding phase, the data is in a meaningful state for the application of data mining techniques.

3.4. Data mining

Data mining, the last phase of the knowledge discovery process, offers automated discovery of previously unknown patterns as well as automated prediction of trends and behaviors.

Data mining tools are most effectively used with data stored in data warehouses or data marts (i.e., after the data has been consolidated, cleansed, and prepared for data mining), even though data mining tools usually also support data stored in flat files or databases.

The next chapter provides more detail about the data mining process. For more information on the knowledge discovery lifecycle please see [2] as well as [12].

4. Data mining lifecycle

The main phases of the data mining process are: study definition, discovery, refinement, and prediction. Each one of these stages is described in more detail in the following sections. As part of the description of the study definition phase, a very brief overview of some of the more commonly used data mining algorithms and methodologies is provided.

The data mining steps are iterative in nature, e.g., based on analyzing the results, the marketing professional might decide to rerun the discovery process with slightly different input or study definitions.

4.1. Study definition

The study definition phase involves defining a data mining goal, identifying the data to be mined, and specifying the appropriate data mining algorithm and methodology.

Goal definition

The goal of a data mining study can be to predict particular outcomes (classification) or to define groups of related characteristics (clustering). Classification and clustering are two of the most popular approaches to data mining. Following is a brief description:

- **Classification**
  Classification studies classify a set of examples vis a vis a particular business goal. Classification analyzes the business records to determine what values in the respective fields have an important influence on the business goal. Classification creates a model which can then be used for predictive purposes for additional records.
- **Clustering**
  Clustering studies identify the kind of data that tends to occur together with other data. Data within a cluster have similarities, but differ significantly from other data outside the cluster. Clustering can be used to determine the characteristics of customers to whom a particular product or rate plan might appeal.

Data identification

Identifying the relevant data involves at a minimum specification of data sources, sampling size, and identifying a strategy for dealing with missing data and binning for continuous variables.

Algorithm specification

After defining the goal and identifying the data to be mined, the business analyst needs to indicate which data mining algorithm should be applied. Many data mining algorithms and variations of established data mining methodologies have been proposed by academia and research. It is not a goal of this article to give a *comprehensive* overview of data mining algorithms. We do, however, present some highlights of the more established methodologies in order to show how they fit into an enterprise approach for data mining.

- **Neural networks:**
  Neural networks have been modeled according to the human brain. They consist of a set of nodes (modeled after neurons) which are connected to each other (like neurons are connected to each other by synapsis). Typically, a neural network consists of a set of input nodes which receive input signals, a set of output nodes which give the output signals, and a number of intermediate layers. Neural nets have a long history and are well established in the applied field. Reference [13] provides both a historical discussion and a review of important applications.

  Neural networks are used in two stages: during the encoding stage (discovery and refinement) the network is trained to perform a certain task, and during the decoding stage (prediction) the network is used for classification and prediction. In its initial stage, a network has random weights on its connections.

  During the encoding stage, a neural networking expert exposes the net to a training set of input data. For each training instance, the actual output of the network is compared with the correct output; if there is a difference between the correct answer and the actual answer, the weights of the individual nodes and connections of the network are adjusted. This process is repeated until the responses are more or less accurate.

  Once a neural network has been trained, it has very good predictive capabilities. Neural networks handle noise well and are robust with respect to dealing with missing data. A couple of disadvantages of neural networks are that (a) an artificial intelligence background is required to properly train the network, and (b) that it does not provide the business analyst with a theory about how it has learned (i.e., it is a black box approach). More drawbacks of using neural networking technology are the lack of scalability with respect to large number of records and number of input fields, and the amount of time it takes to properly train a network. Neural networks are also prone to overfitting: they become very good at predicting the test data at the expense of accuracy of new data.

- **Decision trees:**

The basic idea behind decision trees is as follows: the objective is to find the most discriminating field vis a vis a stated business goal. The algorithm initially finds the first field and a threshold for that field, splits the field in two and goes on to the next field. Again it finds a threshold, and repeats this process recursively until a correct classification for most records in the data source is found, thus creating a decision tree. Well-known examples of decision tree algorithms are CART (see [3]), XAID/CHAID methods (see [15]), ID3 (see [18]), and its successor C4.5 (see [19]).

Following is an example of how a decision tree algorithm might work. During a classification study of customers who buy a particular product, a business analyst might find out that age is the most discriminating field (e.g., customers below the age of 45 are likely to purchase the product, whereas customers above 45 do not purchase the product). Within the group of customers who are below 45, the next most discriminating field might be income (e.g., customers below a certain income level do not purchase the product). This process is repeated for potentially multiple levels in each branch of the tree until it does not further improve the overall level of accuracy or amount of information.

The biggest advantage of this algorithm and methodology is its conceptual simplicity. It generates straight forward if-then-else rules which most business analysts can follow and interpret (i.e., it alleviates the black box issue of neural networks). Another benefit of decision trees is that the learning phase is much faster than training a neural network.

The simplicity of decision trees comes at the cost of not scaling well to larger data sets from a number of angles. First, decision trees cannot be adapted for incremental learning (i.e., updating a model with additional data cannot be accomplished without re-starting the learning process from scratch). Secondly, performance does not scale linearly or sublinearly with size. Thirdly, the results are hard to visualize beyond very simple examples. Another disadvantage is that data ordering affects the performance. Another shortcoming of decision trees is they cannot uncover rules based on combinations of variables.

- **Genetic algorithms:**

The theory behind genetic algorithms is inspired by biology (particularly by the mechanism of natural selection). Genetic algorithms have evolved from genetic research, evolutionary programming, and evolutionary strategies. The mechanism of natural selection has both advantages and disadvantages, the major disadvantage being the large over-production of individuals. The main advantage of natural selection on the other hand is its robustness which is mostly based on the large number of simultaneous experiments: if there is something to be found, it usually is; genetic algorithms do not get trapped in local optima. [14] describes genetic algorithms as search techniques designed to solve hard optimization problems.

The genetic algorithm designer first needs to (a) devise a good coding of the problem, (b) invent an artificial environment where the solutions can join the battle with each other, and (c) provide a fitness function which is used to objectively rate the results of an experiment as a success or failure. The definition of cross-over operations, mutation operators, and an initial population finishes the definition phase. At that point the computer takes over and plays evolution until a successful solution to the problem has been found.

The main advantage of using a genetic algorithm is akin to the advantage of natural selection in the sense that it is a solid technique which finds a solution if one exists. The large over-production of solutions on the other hand requires a lot of computing power and an expert is required to code the problem as well as the artificial environment necessary for the evolutionary game.

- **Agent Network Technology:**

The Agent Network Technology has been developed for handling business data mining problems. It is based on Bayesian probability (see [20]) and combines ideas from the artificial intelligence community (e.g., rules, semantic networks) with connectionist association methods (e.g., neural networks) to build networks for inductive modeling.

The basic building blocks of these networks are software agents. Agents are ‘free form’ in nature and do not have a predetermined structure or organization. In fact, agents are self organizing in that they dynamically determine their relationships together and change these relationships over time as necessary. Agent networks may be multi-layered and might contain feedback connections.

The agent network is built automatically during the discovery phase of the data mining process; it *is* the data mining model. Agent networks consist of agents, connections between agents, and weights on the connections between agents. An agent is created for each unique input field value read from a data source and for each unique output field value which represents the business goal which the business analyst intends to explore. Agents represent data; connections between agents represent data relationships; weights represent the impact of input field values or conjunctions of input field values on output field values. Weight values are

determined by applying a modified Bayesian probability method.

The agent network is used for evaluating the accuracy of the model against different data sets as well as for prediction purposes. Reference [16] gives a detailed description of the agent network technology including the algorithmic equations which lie at the heart of it. Reference [4] provides more of an overview for the technically inclined business reader.

The main advantages of the agent network technology are: It combines the advantages of best of breed technologies while avoiding the disadvantages of a single approach, it is designed for the business user, it scales well, and it is very fast and accurate.

Having specified goals, data, and data mining algorithm for a particular data mining study concludes the study definition phase. Study definitions are stored in repositories to support enterprise wide development.

4.2. Discovery

After completion of the study definition, the marketing professional initiates the discovery process, which creates a model of the data. Data records are read from data sources according to the study definition, and the information from the data sources is condensed by data mining learning algorithms into models.

Data mining tools use symbolic classifiers to perform inductive reasoning which involves learning by example. Symbolic classifiers operate by evaluating and testing a multitude of hypotheses in an effort to discover which factors or combination of factors have the most influence on the dependent outcome. Data mining is automated, determined by the properties of the records in the database, and generalized into models from specific examples in the data set.

4.3. Refinement

After the discovery phase, the business analyst uses various metrics and graphs to understand the model results. One very important activity during this phase is evaluating the predictive accuracy of the model with respect to other data sets. Using the insights gained by understanding the model and measuring its effectiveness, the business analyst then starts the process of refining the model by making adjustments to the data or the study parameters, and regenerating and re-measuring the effectiveness of the model. This iterative process continues until the analyst is satisfied with the predictive capabilities of the model.

Data mining tools need to support the generation of very easy-to-understand metrics and graphs for the business analyst as well as more complex statistical metrics for the analyst who works together with a statistician to gain more understanding of the data.

4.4. Prediction

The last phase of the iterative data mining process is the prediction phase. During this phase, a model is applied to additional data to gain insight with respect to the goals that were specified for the study.

5. Data mining issues and requirements

This chapter describes data mining functionality, development, and deployment requirements from an enterprise perspective. ‘Enterprise’ in this context means multi-tiered topologies including Intranet and Internet, group development, and mass deployment of data mining applications and services. The deployment phase of data mining applications makes the predictive models and studies available to a larger user community. In the previous chapter of this article we have referred to this step in the lifecycle as prediction.

5.1. Enterprise data mining functionality

All of the data mining functionality important for an individual analyst is also important from an enterprise standpoint. Support for the following functions, however, deserves a closer look regarding issues and requirements for the enterprise:

- **Heterogeneous data sources**
  Enterprise data mining applications are not different from other enterprise applications which require access to multiple heterogeneous data sources. Data mining services are dependent on information services which provide transparent and efficient access to relational database systems, data warehouses, data marts, flat files, object oriented database systems, Web pages, and mainframe hierarchical database systems.
- **Sampling**
  Sampling refers to the capability of selecting a subset of a data source for building a model, evaluating a model, or using a model for prediction. Support for this functionality is critical for enterprise data mining applications for the following two reasons:
  1. Scalability:
     It is impractical to build a data mining model for very large data sets. It is, however, a valid methodology to build a model from one or more significant samples of the data source and then validate the model against multiple other samples of the same or different data sources.

2. Support for data mining lifecycle:
   In order to speed up the data mining lifecycle, the business analysts might first use subsets of data from different data sources before including all or most of the data in the study definition.

For enterprise applications it is a requirement to support different sampling sizes for different heterogeneous data sources. In an ideal architecture, this level of functionality would and should be supported by the data sources directly. Today, however, it is part of the data mining services layer.

- **Model merge**
  Model merge refers to the capability of combining two or more data mining models which have been created independently. This functionality is critical for enterprise data mining applications for the following two reasons:
  1. Scalability:
     Very large models from very large data sets can be built by creating multiple sub-models from disjoint sub-sets of data in parallel, and then using the model merge capability to combine them into one very large model.
  2. Business process support:
     An enterprise might have different independent models for each geographic region, but might also need a single combined model representing the entire nation. Model merge enables this type of enterprise data mining.
- **Incremental modeling**
  Incremental modeling is similar to model merge with the following difference: the process of incremental modeling starts out with a single model and then updates this model with additional data thus creating a single updated model which contains the information of the original model combined with the information of the additional data. This functionality is critical for enterprise data mining applications for the following two reasons:
  1. Scalability:
     An organization might have very large models and might need to periodically update these models with new incremental data. In this scenario, it is very important that the data mining service does not have to recreate the original model from scratch, but rather can start out with the original model and enhance it with additional data.
  2. Business process support:
     An organization which maintains data mining models, requiring periodic data updates, requires this type of functionality.

5.2. Enterprise data mining development

Enterprise data mining development implies support for the data mining lifecycle activities for a globally distributed group of business professionals working on a common data mining problem. Following is a description of the issues and requirements for multi-user development, security, and administration, followed by examples of this type of group development.

- **Multi-user development**
  Effective multi-user development of studies requires a repository for study definitions to support sharing and refinement of studies by multiple business analysts. In order to avoid simultaneous updates of studies, the repository needs to provide concurrency control via a lock and unlock facility. Additional study repository services needed for enterprise development are publish and subscribe facilities. Publishing a study makes it available for other business analysts to use. Subscribing to a study provides notification when a study changes.

- **Security and administration**
  The user authentication and authorization model for data mining applications needs to be aligned with the user model for data source access. Data mining services need to augment the general user authorization model by providing different access levels for updating, deleting, and read-only access to studies. Administration tools for system administrators as well as business analysts need to provide facilities to create, delete, copy, rename, upload, download, lock, and unlock studies.

5.3. Enterprise data mining deployment

Enterprise data mining deployment implies the distribution of predictive models and studies throughout a large user community. Following is a description of the issues of scalability, reliability, and recoverability.

- **Scalability**
  Predictive models are integrated into simple (usually thin client) applications and deployed to a large number of users. This implies that scalability of the prediction engine to thousands of concurrent users is an important requirement. Studies are shared between users for simultaneously executing predictions.

  The prediction engine executing on a server machine needs to scale with processing power (e.g., number of processors, available RAM); reasonable interactive performance (e.g., measured in response time) must be guaranteed. The data mining service implementation needs to optimize networking traffic between the data sources and the data mining server on one hand, and

between the data mining server and the clients on the other hand. High end scalability of the data mining server implies sophisticated scheduling capabilities of data mining jobs.

- **Reliability and recoverability**

  Enterprise deployment of data mining applications implies robustness of data mining services with respect to client failures and server failures.

  Resiliency with respect to client failures includes the following requirement: if a connection goes down, either intentionally or not, the client must have a way to retrieve results later, possibly from a different machine.

  Resiliency with respect to server failures includes the following requirement: if the server goes down during processing, a recovery mechanism needs to be provided which picks up where the server left off before it went down. This can be achieved either through check pointing or some failover facility.

6. Data mining integration

This chapter describes the various integration architectures of data mining services within an enterprise topology. The following four integration configurations are addressed: data warehouse integration, application server integration, thick and thin client integration.

6.1. Data warehouse integration

The architecture of a tightly integrated data mining service within a data warehouse implies that the data mining service is linked in with the data warehousing server. To the data warehousing users, data mining appears as an extension of the data warehouse. In this architecture, the data mining service is a 'plug-in' of the data warehouse. Most database vendors and data warehousing vendors are providing or will provide a mechanism for plugging in this type of service. For an example of a tight integration of a data warehouse with data mining services see [10]. [1] presents a methodology for tightly coupling data mining applications to database systems.

In the data warehouse integration architecture, data mining as well as data source access is done on the server. Data sources are managed by the data warehouse. This has significant scalability implications with respect to networking traffic and storage requirements. There is no need to transfer data from one server to another in order to create models. Data can be fed directly into the data mining service for algorithmic analysis without having to create intermediate files. The data mining service might use the data warehousing storage engine for study repository management in order to simplify system management of study objects and users.

The data warehouse server integrates with the data mining service using the data mining application programming interface (API) which provides full programmatic control of the mining process and its parameters (see [8]).

6.2. Application server integration

For enterprise data mining applications which require access to multiple heterogeneous data sources, this architecture provides the most flexibility. While it is still possible to co-locate the data warehousing server with the data mining server, it is not required. The data mining services are implemented within their own process infrastructure and can be executed on a stand-alone server. The data mining API is published through an object oriented communication infrastructure and other applications might integrate with the data mining service either on the server or the client. Reference [6] provides a description of a server-based data mining implementation.

The data mining server itself needs to provide an extensible architecture with respect to data mining algorithms. Reference [21] describes an example of an extensible data mining system built around a 'plug-in' architecture.

Data source adaptors provide the architectural flexibility and extensibility to deal with specialized data sources, to add new data sources, and to take advantage of specific data mining enhancements of a particular data source implementation.

6.3. Thick client integration

The thick client model is suitable for the occasional data mining exploration with small data sources. In this configuration, data mining is accomplished as an extension to assumption driven decision support tools. The data mining service executes on the client machine; the data source is located either locally (i.e., in a flat file, a spreadsheet, or a local database) or on a server (i.e., in a flat file or in a database system which supports ODBC). Models are built locally, and the study repository can be located on a server machine to support sharing of study definitions. Reference [5] provides a description of a thick client-based data mining implementation.

Data mining templates provide support for various data mining methodologies (e.g., clustering template, classification template, etc.). Data mining templates can be customized (e.g., user interface customization, data source customization) to support a team of system analysts or business professionals who all have similar requirements with respect to the data being mined.

Solution templates provide support for various (vertical) application specific activities which include data mining (e.g., churn prediction template, telecom customer segmentation template, cross-selling template, etc.). Solution templates can be customized (e.g., user interface customization, data source customization) to optimally integrate into a solution.

6.4. Thin client integration

In the thin client integration architecture, the data mining service executes on the server. Only templates and mining applets execute on the client machine. This architecture is particularly applicable to and suitable for prediction-only applications which do not require much sophistication on the client side.

A Web browser can be used to view the results of a discovery or prediction process. The data mining service generates HTML reports which can then be viewed via the browser. Downloadable mining applets can be used to execute data mining specific tasks on the thin.

7. Data mining architecture components and distributed information services

This chapter summarizes the various data mining architecture components and elaborates on their relationship to other components of a distributed heterogeneous information services infrastructure.

7.1. Distributed heterogeneous information services infrastructure

Information services generate information by communicating and manipulating data from distributed sources. There are two layers of infrastructures that provide the necessary tools for generating information services: Information services infrastructure and communication infrastructure.

7.2. Data mining services

An implementation of data mining services provides a set of data mining application programming interfaces (APIs) to applications and applets. Data mining services are part of the distributed heterogeneous information services architecture; they leverage services and infrastructure from layers below:

- Data mining services utilize data sources to implement sampling, data access, statistics gathering, data preparation, etc.
- Data mining services utilize information services to access various data sources.
- Data mining services utilize distributed heterogeneous information services to access distributed and heterogeneous data sources.

7.3. Study definition services

An implementation of study definition services provides a set of study repository services to the data mining services layer. Study definition services are part of the information services infrastructure; they leverage services from the layer below, i.e., they utilize data sources as the storage engine underlying the study repository.

System administration tools and services utilize study definition services to provide system management utilities which allow for the management of studies and users in an enterprise context.

8. Conclusion

Let us conclude with a brief analysis of some of the current trends relevant to enterprise usage of data mining tools and methodologies:

- **Standardization**
  We anticipate an increase in standardization efforts of data mining application programming interfaces.
- **Benchmarks**
  We have only scratched the surface in our description of different data mining algorithms and methodologies. Even though there is some consolidation of core data mining methodologies going on, variations of existing algorithms as well as new ones are plentiful. This makes it challenging to compare the various methods with respect to accuracy and performance. The specification of generic as well as industry specific benchmarks will be necessary to alleviate this problem.
- **Enterprise scalability**
  The push towards more scalability will continue into the foreseeable future. Database and data warehousing vendors will update their core database offerings to better enable data mining. They will add better sampling support, better and more accurate support for statistical and meta data information, and their respective ‘plug-in’ architectures will become more mature such that data mining functionality can be implemented as an extension to the core database and data warehousing servers. Internet and intranet architectures will be fully embraced
- **Self-sufficiency**
  The trend towards more self-sufficiency with respect to data mining projects will continue. Organizations will demand data mining solutions which can be maintained in-house.
- **Active research**

Data mining will stay a very active area within the research community which will continue to explore data mining algorithms and methodologies, sophisticated visualization techniques, and new user interface paradigms which satisfy the ease-of-use requirements.

Acknowledgments and Disclaimer

This article reflects the work of many people. I like to thank all members of the DataMind product development team for pioneering and developing these concepts and architectures. I also wish to thank AJ Brown and Ram Srinivasan for sharing their insights into datamining requirements which had a major impact on the enterprise datamining architecture described in this article.

The contents of this article reflect the view of the author only. It does not necessarily represent the product direction of DataMind Corporation, nor does it imply that the current product releases of Professional Edition, or DataCruncher conform to the description in this article.

References

[1] Agrawal R., Shim K., 'Developing Tightly-Coupled Data Mining Applications on a Relational Database System', Proceedings of Second International Conference on Knowledge Discovery & Data Mining, edited by Evangelos Simoudis, Jiawei Han, & Usama Fayyad, Portland, Oregon, August 1996.

[2] Adriaans P., Zantinge D., 'Data Mining', Addison-Wesley, 1996.

[3] Breiman L., Friedman J.H., Olshen R.A., Stone C.T., 'Classification and Regression Trees', Belmont, California: Wadsworth, 1984.

[4] DataMind Technology White Paper, 'Agent Network Technology', http://www.datamindcorp.com, 1996.

[5] DataMind, 'DataMind Professional Edition User Guide', 1997.

[6] DataMind, 'DataMind DataCruncher User Guide', 1997.

[7] DataMind, 'Mining Data to Understand and Predict Telecom Churn', 1997.

[8] DataMind, 'DataMind External API', 1997.

[9] DataMind Technology White Paper, 'A Model for Effective Customer-Oriented Market Plans', http://www.datamindcorp.com, 1997.

[10] Fernandez P., 'Increasing DW "Smarts" with Integrated Data Mining', Proceedings of the DCI Data Warehousing Conference, Phoenix, September 1996.

[11] Feelders A.J., le Loux A.J.F., van't Zand J.V., 'Data Mining for loan evaluation at ABN AMRO: a case study', Proceedings of First International Conference on Knowledge Discovery & Data Mining, edited by Usama M. Fayyad & Ramasamy Uthurusamy, Montreal, Canada, August 1995.

[12] Fayyad U., Piatetsky-Shapiro G., Smyth P., 'Knowledge Discovery and Data Mining: Towards a Unifying Framework', Proceedings of Second International Conference on Knowledge Discovery & Data Mining, edited by Evangelos Simoudis, Jiawei Han, & Usama Fayyad, Portland, Oregon, August 1996.

[13] Hecht-Nielsen, R., 'Neurocomputing', Addison Wesley, Reading, Mass., 1990.

[14] Holland J.H., 'Adaptation in Natural and Artificial Systems', Ann Arbor: University of Michigan Press, 1975.

[15] Kass G.V., 'Significance testing in automatic ineraction detection', Applied Statistics 24 (2): 178-189, 1975.

[16] Pham K.M., 'The NeurOagent: A Neural Multi-agent Approach for Modelling, Distributed Processing and Learning', Intelligent Hybrid Systems, edited by S. Goonatilake and S. Khebbal, John Wiley & Sons Ltd., 1995.

[17] Piatetsky-Shapiro G., Brahman R., Khabaza T., Kloesgen W., Simoudis E., 'An Overview o Issues in Developing Industrial Data Mining and Knowledge Discovery Applications', Proceedings of Second International Conference on Knowledge Discovery & Data Mining, edited by Evangelos Simoudis, Jiawei Han, & Usama Fayyad, Portland, Oregon, August 1996.

[18] Quinlan, J.R., Learning efficient classification procedures. 'Machine learning: an artificial intelligence approach', edited by Michalski R., Carbonell J., Mitchell T., Palo Alto, California, Tioga Press, 1983.

[19] Quinlan, J.R., 'C4.5 Programs for Machine Learning', San Mateo, CA: Morgan Kaufmann, 1993.

[20] Sivia D.S., 'Data Analysis, A Bayesian Tutorial', Clarendon Press Oxford, 1996.

[21] Wrobel S., Wettschereck D., Sommer E., Emde W., 'Extensibility in data mining systems', Proceedings of Second International Conference on Knowledge Discovery & Data Mining, edited by Evangelos Simoudis, Jiawei Han, & Usama Fayyad, Portland, Oregon, August 1996.