View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

DATA MINING APPLICATIONS ON WEB USAGE ANALYSIS & USER PROFILING

M.Sc. Thesis by

Osman Onat ÜNAL, B.Sc.

Department : Management Engineering

Programme: Management Engineering

Supervisor : Dr. Halefşan SÜMEN

SEPTEMBER 2003

İSTANBUL TECHNICAL UNIVERSITY ★ INSTITUTE OF SCIENCE AND TECHNOLOGY

DATA MINING APPLICATIONS ON WEB USAGE ANALYSIS & USER PROFILING

M.Sc. Thesis by

Osman Onat ÜNAL, B.Sc.

(507001133)

Date of submission: 5 May 2003

Date of defence examination: 22 September 2003

SUPERVISOR (CHAIRMAN): Inst.Dr. Halefşan SÜMEN

Members of the Examining Committee Prof.Dr. Burç ÜLENGİN (İTÜ.)

Prof.Dr. Nahit SERARSLAN (İTÜ.)

SEPTEMBER 2003

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ FEN BİLİMLERİ ENSTİTÜSÜ

INTERNET KULLANIM ANALİZİ VE KULLANICI BETİMLEME KONULARINDA VERİ MADENCİLİĞİ UYGULAMALARI

YÜKSEK LİSANS TEZİ

İşletme Mühendisi Osman Onat ÜNAL

Tezin Enstitüye Verildiği Tarih : 5 Mayıs 2003

TEZ DANIŞMANI : **Öğr.Gör.Dr. Halefşan SÜMEN**
Diğer Jüri Üyeleri **Prof.Dr. Burç ÜLENGİN (İTÜ.)**
Prof.Dr. Nahit SERARSLAN (İTÜ.)

EYLÜL 2003

FOREWORD

It can be foreseen that the data mining technology will be one of the key functions of the knowledge discovery process in business. It has wide application areas in many industries and sciences including bioinformatics, geographical information systems, and business decision support. The rapid growth of internet and computer technologies resulted in mass accumulation of data. Understanding the interaction between websites and customers became extremely important. In this point the aim of this thesis is to understand Data Mining Technology and apply its techniques to web usage analysis and user profiling.

I wish to express my deep gratitude to my supervisor and professor Halefşan Sümen. I thank him for his continuous encouragement, his confidence and support that did not last for years, and for sharing with me his vision on MIS and data mining.

I am very thankful to my professor Bertan Badur, Instructor in Boğaziçi University Management Information Systems Department, for his valuable comments and advice. His observations and suggestions for improvement were very assuring and helped me to overcome the case.

I would like to send my appreciations to my colleagues in Boğaziçi University Management Information Systems Department and friends who encouraged and helped me to finish this study.

July, 2003 Osman Onat ÜNAL

TABLE OF CONTENTS

ABBREVIATIONS

AI	: Artificial intelligence
APR	: Annual Percentage Rate
CRISP-DM	: Cross Industry Standard Process for Data Mining
CRM	: Customer Relationship Management
HTML	: Hyper Text Mark Up Language
IIS	: Internet Information Server
KDD	: Knowledge Discovery in Databases
MB	: Market Basket
MS	: Microsoft
NPV	: Net Present Value
OLAP	: Online Analytical Processing
OLTP	: Online Transaction Processing
PE	: Processing Element
SQL	: Structured Query Language

LIST OF TABLES

PageNo.

LIST OF FIGURES

Page No.

DATA MINING APPLICATIONS ON WEB USAGE ANALYSIS & USER PROFILING

ABSTRACT

Recent advancements in computer processing and storage technologies enabled business enterprises to capture their transactional data easily. The new problem is the analysis and interpretation of this data which can not be done merely with database technologies or classical statistical techniques. Data mining and data warehousing technologies solve the problem of storage and usage of mass data in order to maintain knowledge. Data mining technology can handle huge amounts of data and reveal hidden relations and patterns by automatic and semi-automatic discovery techniques.

Enterprises are trying to move beyond the basics of business transactions into a deeper level of understanding of what is occurring at their website. This data if combined with customer demographic information has the potential of revealing the secrets behind the implicit communication between the user and the website. It is obvious that defining this interaction provides a significant competitive advantage to the enterprises in global markets. Usage of this knowledge can provide better decisions and result with better marketing, better retailing and better profits. On the other hand customer can benefit from more qualified and customized service.

Although the main focus is the web user and website interaction, this thesis gives a summary of data mining technology, its functionalities and applications. OLAP technology and data warehouses are also introduced as the key concepts in data mining. The usage of data mining on the internet and the decisions based on the Internet usage data are introduced.

In the application section a web retailer's transactional data is used for analyzing customer profiles and customer shopping patterns. The data includes web transaction details and customer demographic information. We tried to extract the patterns within the data in order to support business decisions such as user profiling and

customer segmentation. We used a common database server Microsoft SQL Server to maintain the data. Microsoft Analysis Server is used as the OLAP server. Microsoft Analysis Server also is used as a data mining tool for classification and clustering. SPSS Answer Tree is another tool used for classification. Dbminer (a third party tool for Microsoft Analysis Server), office automations, statistical packages are some of the other software tools used in the analysis.

The results are summarized with the output screens of the software used. The results are evaluated by discussing on the possible decisions that can be supported by the analysis. In the conclusion section the effective usage of data mining technology and its potential on prevailing competitive advantage are discussed.

INTERNET KULLANIM ANALİZİ VE KULLANICI BETİMLEME KONULARINDA VERİ MADENCİLİĞİ UYGULAMALARI

ÖZET

Bilgisayar işlemci yongaları ve depolama teknolojilerindeki son gelişmeler ticari kurumların işlemsel seviyedeki verilerini kolaylıkla tutmalarını sağlamıştır. Yeni problem, sadece veritabanı teknolojileri ve klasik istatistiksel teknikler kullanılarak çözülemeyecek olan, verinin analizi ve yorumlanmasıdır. Veri madenciliği ve veri ambarı teknolojileri, kitlesel veriyi bilgi elde etmek amacıyla depolama ve kullanma problemlerine çözüm getirir. Veri madenciliği teknolojileri büyük miktarlardaki veriyle kullanılabilmesinin yanısıra gizli ilişkileri ve kalıpları otomatik ve yarı otomatik keşif teknikleriyle açığa çıkarabilir.

Kurumlar işlem seviyelerindeki verilerini, internet sayfalarında ne olup bittiğini anlamaya yönelik, daha ileri çıkarım seviyelerinde yorumlamaya çalışmaktadırlar.Bu veriler, müşteri demografik bilgileriyle birleştirildiğinde, kullanıcı ve internet sayfası arasındaki örtük iletişimin ardındaki gizleri açığa çıkarma potansiyeline sahiptir. Bu durumun küresel pazarlarda kurumlara önemli bir rekabet avantajı sağlayacağı açıktır. Bu bilginin kullanımı daha iyi kararlara olanak sağlayacak ve daha iyi pazarlama, parakendecilik ve karlılıkla sonuçlanacaktır. Öte yandan müşteriler daha kaliteli ve kişiselleştirilmiş hizmetten yararlanabilecektir.

Tezin odak noktası internet sayfası ve kullanıcı arasındaki etkileşim olmakla birlikte, veri madenciliği teknolojisinin fonksiyonları ve uygulamaları konusunda özet bilgiler de içermektedir. OLAP teknolojilerine ve veri ambarlarına da veri madenciliğinin anahtar kavramları olarak değinilmiştir. Veri madenciliğinin internette kullanımı ve internet sitesi kullanımına dayalı kararlar açıklanmıştır.

Uygulama kısmında müşteri ve alışveriş kalıpları analizi için bir internet parakendecisinin işlemsel verileri kullanılmıştır. Veriler internet işlemsel seviye ve sipariş detaylarına ek olarak müşteri demografik bilgilerini de içermektedir. Müşteri segmentasyonu ve kullanıcı betimleme gibi konulardaki kurumsal kararları

desteklemek amacıyla veri içerisindeki kalıplar çıkarılmaya çalışılmıştır. Verilerin saklanmasında ve ulaşılmasında yaygın bir veritabanı sunucusu olan Microsft SQL Server kullanılmıştır. OLAP sunucusu olarak Microsoft Analysis Server kullanılmıştır. Microsoft Analysis Server veri madenciliği aracı olarak sınıflandırma ve öbeklendirme amacıyla da kullanılmıştır. Sınıflandırmada ikincil bir araç olarak SPSS Answer Tree kullanılmıştır. Dbminer (Microsoft Analysis Server üzerine geliştirilmiş bir araç), ofis otomasyonları, istatistik analiz paketleri analizlerde kullanılan diğer yazılım araçlarından bazılarıdır.

Sonuçlar kullanılan yazılımların çıktı ekran görüntüleri de sunularak özetlenmiştir. Sonuçlar analiz tarafından desteklenebilecek kararlar üzerinde tartışılarak değerlendirilmiştir. Sonuç bölümünde veri madenciliği teknolojisinin etkin kullanımı ve rekabet avantajını sürdürme potansiyeli tartışılmıştır.

1. INTRODUCTION

As the internet matures as a new medium for business activities through e-commerce, enterprises try to understand the interaction between their web sites and customers. The transaction data files are the only way to evaluate this interaction. The exponentially growing transaction databases on the internet can not be explored and interpreted without automatic and semi automatic discovery technologies. Data mining technology is a must for this new challenge on the race of competitive world where buyer is the king.

The main functionality of data mining is to extract hidden patterns within mass data. Statistics, machine learning, database technologies and computer science are the base disciplines underlying data mining. Data mining on the internet is a wide application area of data mining. But it has many challenging issues and unclarified problems since it has not been matured yet by the enterprises. The theoretic framework of this thesis aims to form an accumulation to understand and apply data mining technology. The thesis was motivated by identification of the interaction between user and the website in e-commerce in order to take the competitive advantage. The outline of the thesis is as follows.

The Second Chapter of the thesis aims to identify the data mining technology and why it was evolved. Emergent forces that drive enterprises to use data mining technology are summarized. From a database perspective, the characteristics of the most common data environments are explained.

The Third Chapter deals with the phases of applying a data mining technology through data mining methodologies. An industry standard methodology CRISP-DM 1.0 is also introduced. Business understanding on the need of data mining is explained.

In the Fourth Chapter the most advanced data repository for data mining; data warehouse, and its descriptive analysis tool, online analytical processing (OLAP) technology are summarized.

The Fifth Chapter summarizes the most common functionalities of data mining. In other words, this chapter summarizes what kind of analysis can be made by data mining to answer business questions.

The Sixth Chapter summarizes the main characteristics of the algorithms used in data mining technology. Strengths and weaknesses of the given algorithms are identified.

The Seventh Chapter aims to give examples of industry applications of data mining. This chapter introduces how the industry specific business problems are being handled with data mining technology.

In the Eighth Chapter, the strategic importance of web warehouses is emphasized. Application areas of data mining on the internet are summarized. The main focus area of the application chapter, decisions based upon click stream analysis are detailed.

In the application chapter, a real click stream dataset provided from KDD CUP 2000[1] contest is used for the analysis. A warehouse is built and possible analyses are made in order to profile customers by the questions identified. The results are summarized with the output screens of the software used. The results are evaluated by discussing on the possible decisions that can be supported by the analysis.

In the conclusion chapter the effective usage of data mining technology and its potential on prevailing competitive advantage are discussed.

---

[1] A competition within KDD-2000 (The Sixth ACM SIGKDD International Conference on Knowledge Discovery and DataMining), Boston, MA, August 2000. The data is available for download from www.bluemartini.com/kddcup2000

2. DATA MINING AND DATA ENVIRONMENTS

2.1 Definitions

Data mining is a new discipline lying at the interface of statistics, data base technology, pattern recognition, and machine learning, and concerned with secondary analysis of large data bases in order to find previously unsuspected relationships, which are of interest of value to their owners. **(Hand, 1998.)**

Data mining is the semi-automatic discovery of patterns, associations, changes, anomalies, rules, and statistically significant structures and events in data.**(Grossman and others, 1998)**

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and summarize the data in novel ways that are both understandable and useful to the owner. (**Hand, 2001)**

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases. **(Han and Kamber, 2001)**

From definitions above we can conclude the major characteristics of the data mining technology.

Observational Data:The third definition made by Hand stress on “observational data”, as opposed to “experimental data.” Data mining typically deals with data that have already been collected for some purpose other than the data mining analysis (for example, they may have been collected in order to maintain an up-to-date record of all the transactions in a bank). This means that the objectives of the data mining exercise play no role in the data collection strategy. This is one way in which data mining differs from much of statistics, in which data are often collected by using efficient strategies to answer specific questions. For this reason, data mining is often referred to a "secondary" data analysis.

Large Data Sets: The definition also mentions that the data sets examined in data mining are often large. If only small data sets were involved, we would merely be discussing classical exploratory data analysis as practiced by statisticians. When we are faced with large bodies of data, new problems arise. Some of these relate to housekeeping issues of how to store or access the data, but others relate to more fundamental issues, such as how to determine the representativeness of the data, how to analyze the data in a reasonable period of time, and how to decide whether an apparent relationship is merely a chance occurrence not reflecting any underlying reality. Often the available data comprise only a sample from the complete population (or, perhaps, from a hypothetical super population); the aim may be to generalize from the sample to the population. For example, we might wish to predict how future customers are likely to behave or to determine the properties of protein structures that we have not yet seen. Such generalizations may not be achievable through standard statistical approaches because often the data are not (classical statistical) "random samples," but rather "convenience" or "opportunity" samples. Sometimes we may want to summarize or compress a very large data set in such a way that the result is more comprehensible, without any notion of generalization. this issue would arise, for example, if we had complete census data for a particular country or a database recording millions of individual retail transactions.

Understandable Results: The relationships and structures found within a set of data must, of course, be novel. Clearly, novelty must be measured relative to the user's prior knowledge. Unfortunately few data mining algorithms take into account a user's prior knowledge.**(Hand and others, 2001)**

Misnomer in the term “Data Mining”: Han emphasizes the misnomer made in the term “data mining”. The term mining is used after the valuable material being extracted. In the term “data mining” it is vice versa. Data is not mined; we have large amounts of data waiting to be interpreted. **(Han and Kamber, 2001)**

- Gold mining: extracting gold from sand
- Coal Mining: extracting coal from rocks
- Data Mining: extracting knowledge from data.

There are other terms that refer to similar meanings to data mining.

- Knowledge mining from databases
- Knowledge extraction.
- Data/pattern analysis.
- Data archaeology
- Data dredging

Data mining is often set in the broader context of knowledge discovery in databases, or KDD. This term originated in the artificial intelligence (AI) research field. It is often used as a synonym for the term "data mining". Knowledge discovery in databases (KDD) is clearer and more informative than the term data mining despite it is less popular.

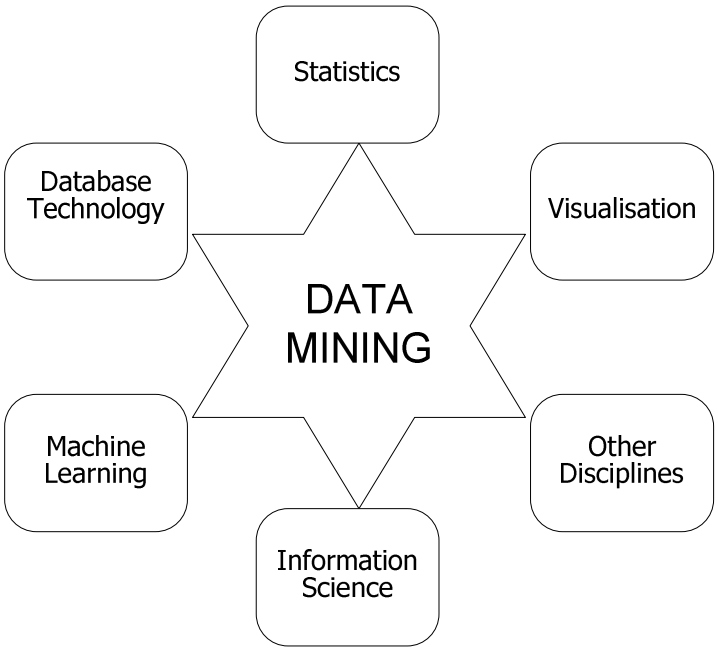


Figure 2-1 Data Mining: Confluence of Multiple Disciplines (Han and Kamber, 2001)

2.2 The Evolution of Data Mining

Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining). For instance, the early development of data collection and database

creation mechanisms served as a prerequisite for later development of effective mechanisms for data storage and retrieval, and query and transaction processing. With numerous database systems offering query and transaction processing as common practice, data analysis and understanding has naturally become the next target. **(Han and Kamber, 2001)**

Table 2-1Emergent forces & Evolution of Data mining (Tiwana, Amrit, 2001)

Emergent Forces	Evolutionary Step	Business Question	Enabling Technology
60s New Products	Data Collection (1960s)	"What was my total revenue in the last five years?"	computers, tapes, disks
70s Low cost manufacturing 80s Total Quality Management	Data Access (1980s)	"What were unit sales in New England last March?"	faster and cheaper computers with more storage, relational databases
90s Customer Relationship Management and one to one marketing.	Data Warehousing and Decision Support	"What were unit sales in New England last March? Drill down to Boston."	faster and cheaper computers with more storage, On-line analytical processing (OLAP), multidimensional databases, data warehouses
2000s Knowledge enabled relationship management and e-business	Data Mining	"What's likely to happen to Boston unit sales next month? Why?"	faster and cheaper computers with more storage, advanced computer algorithms

Since the 1960s, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems. The research and development in database systems since the 1970s has progressed from early hierarchical and network database systems to the development of relational database systems, data modelling tools, and indexing and data organization techniques. In addition, users gained convenient and flexible data access through query languages, user interfaces, optimized query processing, and transaction management. Efficient methods for on-line transaction processing (OLTP), where a query is viewed as a read-only transaction, have contributed substantially to the evolution and wide acceptance of relational technology as a major tool for efficient storage, retrieval, and management of large amounts of data.

Database technology since the mid-1980s has been characterized by the popular adoption of relational technology and an upsurge of research and development activities on new and powerful database systems. These employ advanced data models such as extended-relational, object-oriented, object-relational, arid deductive models. Application-oriented database systems, including spatial, temporal, multimedia, active, and scientific databases, knowledge bases, and office information bases, have flourished. Issues related to the distribution, diversification, and sharing of data have been studied extensively. Heterogeneous database systems and Internet-based global information systems such as the World Wide Web (WWW) have also emerged and play a vital role in the information industry.

The steady and amazing progress of computer hardware technology in the past three decades has led to large supplies of powerful and affordable computers, data collection equipment, and storage media. This technology provides a great boost to the database and information industry, and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis.**(Grossman and others,1998)**

The evolution of data mining and emergent forces are summarized in Table 2-1.

2.3 Data Environments For Data Mining

In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World Wide Web. Advanced database systems include object-oriented and object-relational databases, and specific application-oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases. The challenges and techniques of mining may differ for each of the repository systems.

2.3.1 Relational Databases

A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large set of tuples (records or rows). Each tuple in a relational table represents an object identified by a unique key and described by a set of attribute values. A

semantic data model, such as an entity-relationship (ER) data model, which models the database as a set of entities and their relationships, is often constructed for relational databases.

Relational data can be accessed by database queries written in a relational query language, such as SQL, or with the assistance of graphical user interfaces. In the latter, the user may employ a menu, for example, to specify attributes to be included in the query, and the constraints on these attributes. A given query is transformed into a set of relational operations, such as join, selection, and projection, and is then optimized for efficient processing. A query allows retrieval of specified subsets of the data. **(Han and Kamber,2001)**

Relational databases are one of the most popularly available and rich information repositories, and thus they are a major data form in the area of data mining.

2.3.2 Data Warehouses & Data Marts

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. **(Han and Kamber, 2001)**

Data warehouses are constructed via a process of data cleaning, data transformation, data integration, data loading, and periodic data refreshing.

In order to facilitate decision making, the data in a data warehouse are organized around major subjects, such as customer, item, supplier, and activity. The data are stored to provide information from a historical perspective (such as from the past 5-10 years) and are typically summarized. For example, rather than storing the details of each sales transaction, the data warehouse may store a summary of the transactions per item type for each store or, summarized to a higher level, for each sales region.

A data warehouse is usually modelled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data.

A data warehouse collects information about subjects that span an entire organization, and thus its scope is enterprise-wide.

Data mart, is a department subset of a data warehouse. It focuses on selected subjects, and thus its scope is department-wide.

By providing multidimensional data views and the precomputation of summarized data, data warehouse systems are well suited for On-Line Analytical Processing, or. OLAP operations make use of background knowledge regarding the domain of the data being studied in order to allow the presentation of data at different levels of abstraction

2.3.3 Transactional Databases

In general, a transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number, and a list of the items making up the transaction (such as items purchased in a store). The transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the sales person and of the branch at which the sale occurred, and so on.

2.3.4 Advanced Database Systems and Advanced Applications

Relational database systems have been widely used in business applications. With the advances of database technology, various kinds of advanced database systems have emerged and are undergoing development to address the requirements of new database applications.

The new database applications include handling spatial data (such as maps), engineering design data (such as the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), time-related data (such as historical records or stock exchange data), and the World Wide Web. These applications require efficient data structures and scalable methods for handling complex object structures, variable-length records, semistructured or unstructured data, text and multimedia data, and database schemas with complex structures and dynamic changes. **(Han and Kamber, 2001)**

In response to these needs, advanced database systems and specific application-oriented database systems have been developed. These include object-oriented and object-relational database systems, spatial database systems, temporal and time-series database systems, text and multimedia database systems, heterogeneous and legacy database systems, and Web-based global information systems.

While such databases or information repositories require sophisticated facilities to efficiently store, retrieve, and update large amounts of complex data, they also provide fertile grounds and raise many challenging research and implementation issues for data mining. Leading types of advanced databases and short descriptions are given below.**(Han and Kamber, 2001)**

- Object Oriented Databases: Object-oriented databases are based on the object-oriented programming paradigm.
- Object Relational Databases: Object-relational databases are constructed based on an object-relational data model.
- Spatial Databases: Spatial databases contain spatial-related information.
- Temporal Databases and Time-Series Databases: Temporal databases and time-series databases both store time-related data.
- Text Databases and Multimedia Database: Text databases are databases that contain word descriptions for objects.
- Heterogeneous Database and Legacy Database: A heterogeneous database consists of a set of interconnected, autonomous component databases.
- The World Wide Web: Information repositories used on the internet.

3. PHASES OF A DATAMINING PROJECT

The multidisciplinary structure of data mining and the variety of tasks and procedures in different application areas are leading problems in the way of setting an industry standard methodology. A standard application methodology can make the application of the technology less costly, more reliable, more manageable and faster. A standard methodology will open the way to developers to integrate their data mining solution to fit the overall methodology. A methodology will also make data mining technology more adaptable and understandable.**(Wirth and others, 2001)**

The CRISP-DM (Cross Industry Standard Process for Data Mining) project addressed parts of these problems by defining a process model which provides a framework for carrying out data. The CRISP-DM process model is being developed by a consortium of leading data mining users and suppliers: DaimlerChrysler AG, SPSS, NCR, and OHRA. The project was partly sponsored by the European Commission under the ESPRIT program (Project number 24959) mining projects which is independent of both the industry sector and the technology used.

The CRISP-DM reference model for data mining provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks, and their outputs.

The life cycle of a data mining project is broken down in six phases which are shown in Figure 3-1 . The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase which phase, or which particular task of a phase, has to be performed next. Phases of the CRISP-DM methodology can be overviewed in Table A.1 in Appendix A.

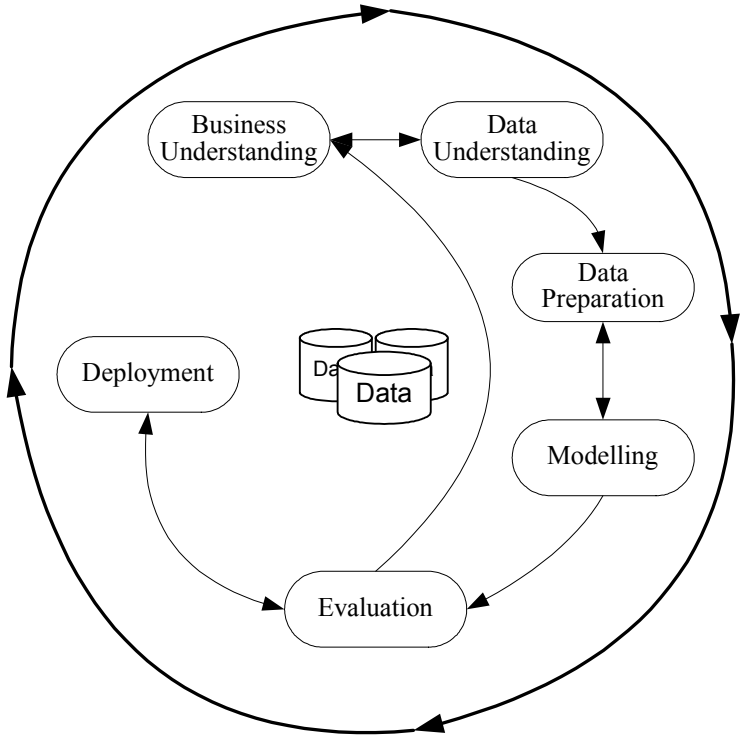


Figure 3-1 Phases of the CRISP-DM Reference Model (Thomas and others, 1996)

3.1 Business Understanding of Data Mining Goals

This initial phase focuses on understanding the project objectives and requirements from a business perspective, and then converting this knowledge into a data mining problem definition, and a preliminary project plan designed to achieve the objectives.

The first and most important step in any targeting-model project is to establish a clear goal and develop a process to achieve that goal. In defining the goal, you must first decide what you are trying to measure or predict. Targeting models generally fall into two categories, predictive and descriptive. Predictive models calculate some value that represents future activity. It can be a continuous value, like a purchase amount or balance, or a probability of likelihood for an action, such as response to an offer or default on a loan. A descriptive model is just as it sounds: It creates rules that are used to group subjects into descriptive categories. Some of the common analytic goals used today in marketing, risk, and customer relationship management is listed below. **(Feeders an others, 2000)**

3.1.1 Profile Analysis

An in-depth knowledge of your customers and prospects is essential to stay competitive in today's marketplace. Some of the benefits include improved targeting and product development. Profile analysis is an excellent way to get to know your customers or prospects. It involves measuring common characteristics within a population of interest. Demographics such as average age, gender (percent male), marital status (percent married, percent single, etc.), and average length of residence are typically included in a profile analysis. Other measures may be more business specific, such as age of customer relationship or average risk level. Others may cover a fixed time period and measure average dollars sales, average number of sales, or average net profits. Profiles are most useful when used within segments of the population of interest.

3.1.2 Segmentation

Targeting models are designed to improve the efficiency of actions based on marketing and/or risk. But before targeting models are developed, it is important to get a good understanding of your current customer base. Profile analysis is an effective technique for learning about your customers.

A common use of segmentation analysis is to segment customers by profitability and market potential. For example, a retail business divides its customer base into segments that describe their buying behavior in relation to their total buying behavior at all retail stores. Through this a retailer can assess which customers have the most potential. This is often called "Share of Wallet" analysis.

A profile analysis performed on a loan or credit card portfolio might be segmented into a two-dimensional matrix of risk and balances. This would provide a visual tool for assessing the different segments of the customer database for possible marketing and/or risk actions. For example, if one segment has high balances and high risk, you may want to increase the Annual Percentage Rate (APR). For low-risk segments, you may want to lower the APR in hopes of retaining or attracting balances of lower-risk customers.

3.1.3 Response

A response model is usually the first type of targeting model that a company seeks to develop. If no targeting has been done in the past, a response model can provide a huge boost to the efficiency of a marketing campaign by increasing responses and/or reducing mail expenses. The goal is to predict who will be responsive to an offer for a product or service. It can be based on past behavior of a similar population or some logical substitute.

A response can be received in several ways, depending on the offer channel. A mail offer can direct the responder to reply by mail, phone, or Internet. When compiling the results, it is important to monitor the response channel and manage duplicates. It is not unusual for a responder to mail a response and then respond by phone or Internet a few days later. There are even situations in which a company may receive more than one mail response from the same person. This is especially common if a prospect receives multiple or follow-up offers for the same product or services that are spaced several weeks apart. It is important to establish some rules for dealing with multiple responses in model development.

3.1.4 Risk

A phone offer has the benefit of instant results. A response can be measured instantly. But a nonresponse can be the result of several actions: The prospect said "no," or the prospect did not answer, or the phone number was incorrect.

Many companies are combining channels in an effort to improve service and save money. The Internet is an excellent channel for providing information and customer service. In the past, a direct mail offer had to contain all the information about the product or service. This mail piece could end up being quite expensive. Now, many companies are using a postcard or an inexpensive mail piece to direct people to a Web site. Once the customer is on the Web site, the company has a variety of available options to market products or services at a fraction of the cost of direct mail.

Approval or risk models are unique to certain industries that assume the potential for loss when offering a product or service. The most well-known types of risk occur in the banking and insurance industries.

Banks assume a financial risk when they grant loans. In general, these risk models attempt to predict the probability that a prospect will default or fail to pay back the borrowed amount. Many types of loans, such as mortgages or car loans, are secured. In this situation, the bank holds the title to the home or automobile for security. The risk is limited to the loan amount minus resale value of the home or car. Unsecured loans are loans for which the bank holds no security. The most common type of unsecured loan is the credit card. While predictive models are used for all types of loans, they are used extensively for credit cards. Some banks prefer to develop their own risk models. Others banks purchase standard or custom risk scores from any of the several companies that specialize in risk score development.

For the insurance industry, the risk is that of a customer filing a claim. The basic concept of insurance is to pool risk. Insurance companies have decades of experience in managing risk. Life, auto, health, accident, casualty, and liability are all types of insurance that use risk models to manage pricing and reserves. Due to heavy government regulation of pricing in the insurance industry, managing risk is a critical task for insurance companies to maintain profitability.

Many other industries incur risk by offering a product or service with the promise of future payment. This category includes telecommunications companies, energy providers, retailers, and many others. The type of risk is similar to that of the banking industry in that it reflects the probability of a customer defaulting on the payment for a good or service.

The risk of fraud is another area of concern for many companies but especially banks and insurance companies. If a credit card is lost or stolen, banks generally assume liability and absorb a portion of the charged amounts as a loss. Fraud detection models are assisting banks in reducing losses by learning the typical spending behavior of their customers. If a customer's spending habits change drastically, the approval process is halted or monitored until the situation can be evaluated.

3.1.5 Activation

Activation models are models that predict if a prospect will become a full-fledged customer. These models are most applicable in the financial services industry. For example, for a credit card prospect to become an active customer, the prospect must respond, be approved, and use the account. If the customer never uses the account, he

or she actually ends up costing the bank more than a nonresponder. Most credit card banks offer incentives such as low-rate purchases or balance transfers to motivate new customers to activate. An insurance prospect can be viewed in much the same way. A prospect can respond and be approved, but if he or she does not pay the initial premium, the policy is never activated.

There are two ways to build an activation model. One method is to build a model that predicts response and a second model that predicts activation given response. The final probability of activation from the initial offer is the product of these two models. A second method is to use one-step modeling. This method predicts the probability of activation without separating the different phases.

3.1.6 Cross-Sell and Up-Sell

Cross-sell models are used to predict the probability or value of a current customer buying a different product or service from the same company (cross-sell). Up-sell models predict the probability or value of a customer buying more of the same products or services.

As mentioned earlier, selling to current customers is quickly replacing new customer acquisition as one of the easiest way to increase profits. Testing offer sequences can help determine what and when to make the next offer. This allows companies to carefully manage offers to avoid over-soliciting and possibly alienating their customers.

3.1.7 Attrition

Attrition or churn is a growing problem in many industries. It is characterized by the act of customers switching companies, usually to take advantage of "a better deal." For years, credit card banks have lured customers from their competitors using low interest rates. Telecommunications companies continue to use strategic marketing tactics to lure customers away from their competitors. And a number of other industries spend a considerable amount of effort trying to retain customers and steal new ones from their competitors.

Over the last few years, the market for new credit card customers has shrunk considerably. This now means that credit card banks are forced to increase their

customer base primarily by luring customers from other providers. Their tactic has been to offer low introductory interest rates for anywhere from three months to one year or more on either new purchases and/or balances transferred from another provider. Their hope is that customers will keep their balances with the bank after the interest converts to the normal rate. Many customers, though, are becoming quite adept at keeping their interest rates low by moving balances from one card to another near the time the rate returns to normal.

These activities introduce several modeling opportunities. One type of model predicts the act of reducing or ending the use of a product or service after art account has been activated. Attrition is defined as a decrease in the use of a product or service. For credit cards, attrition is the decrease in balances on which interest is being earned. Churn is defined as the closing of one account in conjunction with the opening of another account for the same product or service, usually at a reduced cost to the consumer. This is a major problem in the telecommunications industry.

3.1.8 Net Present Value

A net present value (NPV) model attempts to predict the overall profitability of a product for a predetermined length of time. The value is often calculated over a certain number of years and discounted to today's dollars. Although there are some standard methods for calculating net present value, many variations exist across products and industries.

3.1.9 Lifetime Value

A lifetime value model attempts to predict the overall profitability of a customer (person or business) for a predetermined length of tune. Similar to the net present value, it is calculated over a certain number of years and discounted to today's dollars. The methods for calculating lifetime also vary across products and industries.

As markets shrink and competition increases, companies are looking for opportunities to profit from their existing customer base. As a result, many companies are expanding their product and/or service offerings in an effort to cross-sell or up-sell their existing customers. This approach is creating the need for a model that goes beyond the net present value of a product to one that defines the lifetime value of a customer or a customer lifetime value (LTV) model.

3.2 Data Understanding

The data understanding phase starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses for hidden information. There is a close link between Business Understanding and Data Understanding. The formulation of the data mining problem and the project plan require at least some understanding of the available data.

3.3 Selecting Data for Modeling

Data for modeling can be generated from number of sources. Those sources fall into one of two categories: internal or external. Internal sources are those that area generated through company activity such as customer records, web site, mail tapes from mail or ail campaigns, or databases and/or data warehouses that are specifically designed to house company data. The data can be selected from internal resources like;

- Customer Databases: A customer database is typically designed with one record per customer.
- Transaction database: The transaction database contains records of customer activity
- Order/Offer/Purchase History Database: The offer history database contains details about offers made to prospects, customers, or both.
- Datawarehouse: They can be designed for a special purpose and contain historical data.

External sources consist mainly of list sellers, list sellers are companies that sell lists, and compilers.

3.4 Choosing the Modelling Methodology

After defining a clear goal a modelling algorithm should be employed. In the “Data Mining Algorithms” chapter of the thesis, common data mining algorithms and their characteristics are summarized. Statistical methods can be used such as linear

regression or logistic regression. Nonstatistical or blended methods like neural networks, genetic algorithms, classification trees, and regression trees are also commonly used methods.

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques require specific data formats. There is a close link between Data Preparation and Modeling. Often, one realizes data problems while modeling or one gets ideas for constructing new data.

3.5 Data Preparation

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling tool) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

3.5.1 Sampling

Improvements in computer technology reduced the importance of sampling. Without sampling, many analysis can be done. But it requires more professional software tools and computer hardware. This increases the processing and time costs of the process. Since sampling speeds up the process and generally produces the same results, there is no need to avoid sampling.

3.5.2 Maintaining Data- Quality

Data is rarely 100% “clean”. The quality of data on which decisions are made in the corporate world is often suspected. Data mining is at best as good as the data it is representing. The process of purifying data from the problems listed is called Data Cleaning.

Redundant Data: This type of error refers to duplicate records or impossible accuracies. For instance a customer purchased 100 items from the same product in the same day. This data is suspicious to be redundant.

Incorrect or Inconsistent: This refers to invalid data or inconsistent accuracies within data. For instance to see a meaningless name and surname in the customers list.:

Name: Noname

Surname: Noname

This data is suspicious to be incorrect and very difficult to notice.

Or an address with inconsistent entries:

County: USA

City: Ankara

Typos: The computer gets the data as it is written. We can recover typing errors while we are reading but it needs complex algorithms and a knowledge base for computer to do this. Many databases are case sensitive and even capital letters cause problems. For instance the data in the parenthesis shows different typing errors.(Annkara ,Ankara, ANKARa, anlara)

Stale Data: This refers to dynamically changing data. In other words, the data that could have been changed since it is entered. Address, age is typical examples of stale data. Another important factor in data staleness is that the "state of the world" changes. For example, customer behavior and trends change over a time.

Variance In Defining Terms: If the data is combined from different sources. There may be variances in the definitions of the data fields. For instance assume a data collected from the two different plants producing the same products. The field processing time may be calculated by different procedures and techniques so they can not be compared.

3.5.3 Outliers

Outlier analysis is usually performed with continuous variables. An outlier is a single or low frequency occurrence of the value of variable that is far from the mean as well as the majority of the other values for that variable. Determining whether a value is an outlier or a data error is an art as well as a science. Having an intimate knowledge of your data is the best strength.

The outliers in data can show mis typed values or special customers, or in banking and finance fraudulent activities. Outliers can be automatically detected by various statistical analysis.

3.5.4 Missing Values

As information is gathered and combined, missing values are present in almost every data set. Many software packages ignore records with missing values, which makes them a nuisance. The fact that a value is missing, however, can be predictive It is important to capture that information. Missing values can be substituted by:

Single Value Substitution: Single value substitution is the simplest method for replacing missing values. There are three common choices: mean, median, and mode.

Class mean Substitution: Class mean substitution uses the mean values within subgroups of other variables combinations of variables.

Regression Substitution: Similar to class mean substitution, regression substitution uses the mean with subgroups of other variables.

3.6 Selecting and Transforming Variables

In addition to preparing the data for mining, some additional transformation may be necessary. Using data mining to predict behavior may require new variables that have to be derived from the data. For transaction data on existing customers, RFM variables may be good predictors. RFM stands for recency, frequency and monetary. Recency generally would be some measure of time since the last transaction. Frequency would be the number of transactions in a designated period. And Monetary would be the total transactions within a designated period as well as an average per transaction. These additional variables are necessary to make the data more meaningful to the mining process and provide additional parameters from which the mining software may discover useful relationships.

3.7 Processing and Evaluating the Model

Models should be evaluated from a business perspective based on cost benefit analysis and return on investment. The results of the model may show some

interesting patterns but acting on them may not provide the incremental revenue or cost savings that would justify their use.

One of the simplest ways to evaluate a model is to test the results in the real world. Select a sample from the population to test a prediction of the model and see how well the actual results follow the predicted results. The model may predict the likelihood that a certain segment of the market will respond to a particular promotion. By implementing the promotion on a limited sample and testing the results against the prediction, the model's effectiveness can be measured.

3.8 Implementing and Maintaining the Model

Once working models are available, they can be used to understand customer behavior and customer expectations. They may be incorporated in production systems such as campaign management software for marketing purposes. Campaign management software automates marketing campaigns that are used to target customer segments with specific promotions that are most likely to achieve the desired results. Targeting specific segments in this manner should increase the response rate to the promotional campaign based on the model's predictions. This maximizes marketing efficiency and effectiveness. Profiles developed from the data mining project should identify customers that are most likely to respond to cross-selling or up-selling promotions that will lead to an increase customer lifetime value to the organization.

The customer profiles developed from the project can also be used in a Web environment to classify visitors to the site based on their registration information. Then the site can personalize the content presented to them based on the classification. This will increase the likelihood of converting the visitor to a customer. Personalizing the content of the Web site also helps to differentiate the site from the competition and provide a higher level of customer service.

The object is to use the predictive models to drive marketing efforts that will turn Web site visitors into customers and customers into long term clients.

3.8.1 Evaluation

At this stage in the project you have built one or more models that appear to have high quality, from a data analysis perspective. Before proceeding to final deployment of the model, it is important to more thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered.

At the end of this phase, a decision on the use of the data mining results should be reached.

3.8.2 Deployment

Creation of the model is generally not the end of the project. Usually, the knowledge gained will need to be organized and presented in a way that the customer can use it. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process. In many cases it will be the user, not the data analyst, who will carry out the deployment steps. In any case, it is important to understand up front what actions will need to be carried out in order to actually make use of the created models.

4. DATAWAREHOUSE ARCHITECTURE AND OLAP TECHNOLOGY

Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. A large number of organizations have found that data warehouse systems are valuable tools in today's competitive, fast-evolving world. In the last several years, many firms have spent millions of dollars in building enterprise-wide data warehouses. Many people feel that with competition mounting in every industry, data warehousing is the latest must-have marketing weapon. It is considered ad a way to keep customers by learning more about their needs.

According to W. H. Inmon, a leading architect in the construction of data warehouse systems,

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process"

This short, but comprehensive definition presents the major features of a data warehouse. The four keywords, subject-oriented, integrated, time-variant, and nonvolatile, distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems. Let's take a closer look at each of these key features.

Subject-oriented: A data warehouse is organized around major subjects, such as customer, supplier, product, and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Hence, data warehouses typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

Integrated: A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records. Data cleaning and data integration techniques are applied to ensure

consistency in naming conventions, encoding structures, attribute measures, and so on.

Time-variant: Data are stored to provide information from a historical perspective." Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.

Nonvolatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

In sum, a data warehouse is a semantically consistent data store that serves as a physical implementation of a decision support data model and stores the information on which an enterprise needs to make strategic decisions. A data warehouse is also often viewed as architecture, constructed by integrating data from multiple heterogeneous sources to support structured and/or ad hoc queries, analytical reporting, and decision making.

A more informal definition from the business insight enterprises is given below.

Any of a large variety of computer systems initiatives, whose primary purpose is to extract information out of legacy systems, and make it usable to business people, in the support of their efforts to reduce costs and improve revenues.

(Kimball and Merz, 2000)

Data warehousing is the process of constructing and using data warehouses. The construction of a data warehouse requires data integration, data cleaning, and data consolidation. The utilization of a data warehouse often necessitates a collection of decision support technologies. This allows "knowledge workers" (e.g., managers, analysts, and executives) to use the warehouse to quickly and conveniently obtain an overview of the data, and to make sound decisions based on information in the warehouse. Some authors use the term "data warehousing" to refer only to the process of data warehouse construction, while the term "warehouse DBMS" is used to refer to the management and utilization of data warehouses. Many organizations use the information in their warehouse to support business decision making activities, including :

- increasing customer focus, which includes the analysis of customer buying patterns (such as buying preference, buying time, budget cycles, and appetites for spending);
- repositioning products and managing product portfolios by comparing the performance of sales by quarter, by year, and by geographic regions, in order to fine-tune production strategies
- analyzing operations and looking for sources of profit.
- managing the customer relationships, making environmental corrections, and managing the cost of corporate assets.

Data warehousing is also very useful from the point of view of heterogeneous database integration. Many organizations typically collect diverse kinds of data and maintain large databases from multiple, heterogeneous, autonomous, and distributed information sources. To integrate such data, and provide easy and efficient access to it, is highly desirable, yet challenging. Much effort has been spent in the database industry and research community towards achieving this goal.

(Watson and others, 2002)

Data warehousing provides an interesting alternative to the traditional approach of heterogeneous database integration described above. Rather than using a query-driven approach, data warehousing employs an update-driven approach in which information from multiple, heterogeneous sources is integrated in advance and stored in a warehouse for direct querying and analysis. Unlike online transaction processing databases, data warehouses do not contain the most current information. However, a data warehouse brings high performance to the integrated heterogeneous database system since data are copied, preprocessed, integrated, annotated, summarized, and restructured into one semantic data store. Furthermore, query processing in data warehouses does not interfere with the processing at local sources. Moreover, data warehouses can store and integrate historical information and support complex multidimensional queries.

4.1 Operational Database Systems and Data Warehouses

The major task of on-line operational database systems is to perform on-line transaction and query processing. These systems are called on-line transaction processing (OLTP) systems. They cover most of the day-to-day operations of an organization, such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of the different users. These systems are known as on-line analytical processing (OLAP) systems.

The major distinguishing features between OLTP and OLAP are summarized as follows.

Users and system orientation: An OLTP system is customer-oriented and is used for transaction and query processing by clerks, clients, and information technology professionals. An OLAP system is market-oriented and is used for data analysis by knowledge workers, including managers, executives, and analysts.

Data contents: An OLTP system manages current data that, typically, are too detailed to be easily used for decision making. An OLAP system manages large amounts of historical data, provides facilities for summarization and aggregation, and stores and manages information at different levels of granularity. These features make the data easier to use in informed decision making.

Database design: An OLTP system usually adopts an entity-relationship (ER) data model and an application-oriented database design. An OLAP system typically adopts either a star or snowflake model and a subject-oriented database design.

View: An OLTP system focuses mainly on the current data within an enterprise or department, without referring to historical data or data in different organizations. In contrast, an OLAP system often spans multiple versions of a database schema, due to the evolutionary process of an organization. OLAP systems also deal with information that originates from different organizations, integrating information from many data stores. Because of their huge volume, OLAP data are stored on multiple storage media.

Access patterns: The access patterns of an OLTP system consist mainly of short, atomic transactions. Such a system requires concurrency control and recovery mechanisms. However, accesses to OLAP systems are mostly readonly operations (since most data warehouses store historical rather than up-to-date information), although many could be complex queries.

Other features that distinguish between OLTP and OLAP systems include database size, frequency of operations, and performance metrics. These are summarized in Table 4-1. **(Han and Kamber,2001)**

Table 4-1 Differences Between OLTP and OLAP systems. (Han and Kamber, ,2001)

Feature	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

4.2 OLAP Cubes

A data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.

In general terms, dimensions are the perspectives or entities with respect to which an organization wants to keep records. Figure 4-1 shows Foodmart[2]. We can create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, product, customer, store, and promotion. These dimensions allow the store to keep track of things like monthly sales of items, the branches and locations at which the items were sold.

[2] Foodmart is a sample database in Microsoft Analysis Server. FoodMart is assumed to be a large grocery chain operating in the United States, Mexico, and Canada.

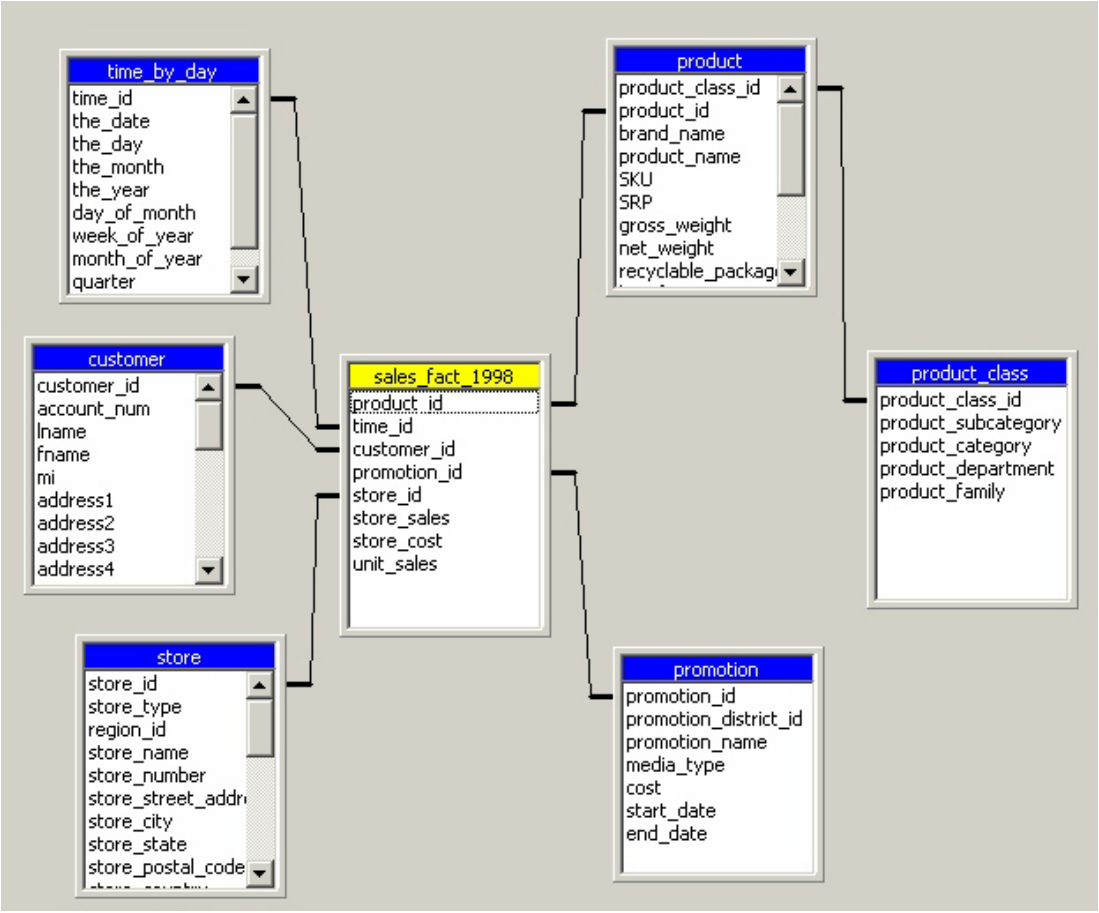


Figure 4-1 Dimension Tables of an OLAP cube from Microsoft Analysis Server

Each dimension may have a table associated with it, called a dimension table, which further describes the dimension. For example, a dimension table for store may contain the attributes store type, store name, store location etc. Dimension tables can be specified by users or experts, or automatically generated and adjusted based on data distributions.

A multidimensional data model is typically organized around a central theme, like sales, for instance. This theme is represented by a fact table. Facts are numerical measures. They can be considered as the quantities by which we want to analyze relationships between dimensions. Examples of facts for a sales data warehouse include store_sales (sales amount in dollars), units_sales (number of units sold), and store_cost. The fact table contains the names of the facts, or measures, as well as keys to each of the related dimension tables.

Although we usually think of cubes as 3-D geometric structures, in data warehousing the data cube is n-dimensional. Conceptually, we may also represent the demographic data of our customers in the form of a 3-D data cube, as in Figure 4-2.

The Figure 4-2 shows the visualization of the foodmart cube according to the selected dimensions and measures.

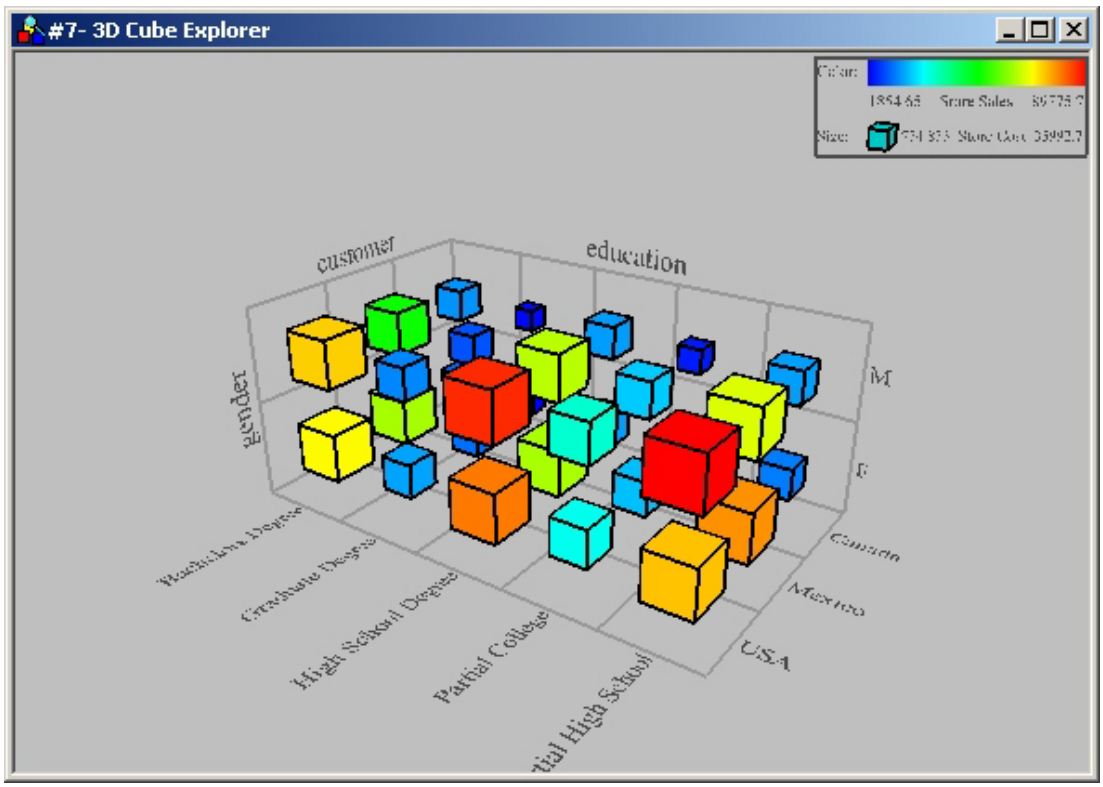


Figure 4-2 3-D Cube Explorer Screenshot from DBminer Software

4.3 Schemas for Multidimensional Databases

The entity-relationship data model is commonly used in the design of relational databases, where a database schema consists of a set of entities and the relationships between them. Such a data model is appropriate for on-line transaction processing. A data warehouse, however, requires a concise, subject-oriented schema that facilitates on-line data analysis.

The most popular data model for a data warehouse is a multidimensional model. Such a model can exist in the form of a star schema, a snowflake schema, or a fact constellation schema.

Star schema: The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (fact table) containing the bulk of the data, with no redundancy, and (2) a set of smaller attendant tables (dimension tables), one for each dimension. In the star schema, each dimension is represented by

only one table, and each table contains a set of attributes. An example of a star scheme view from sales cube in Foodmart Database is illustrated in Figure 4-3.

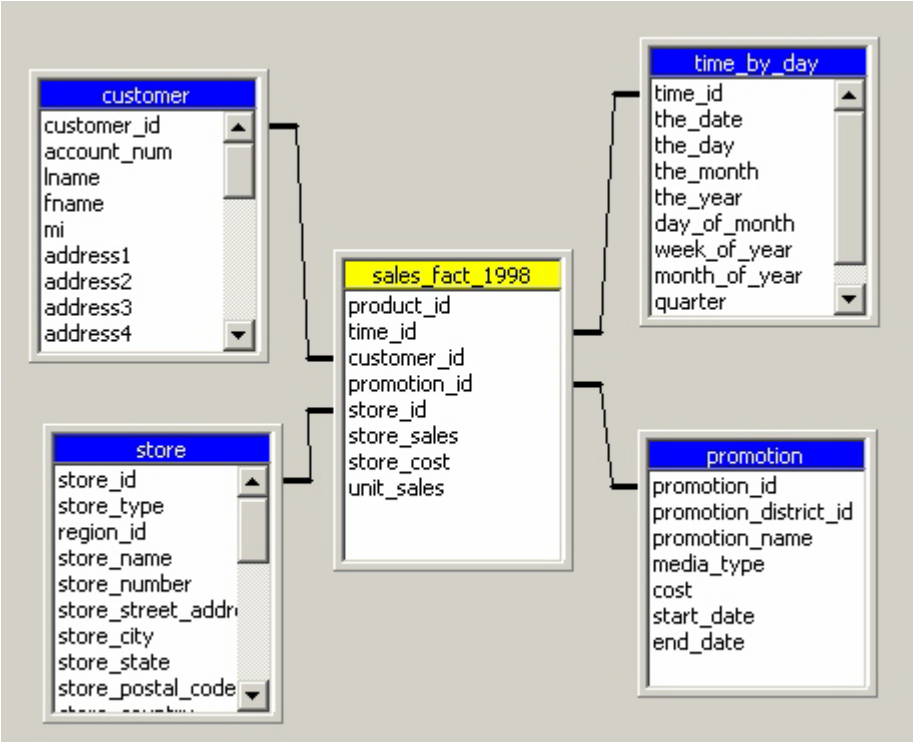


Figure 4-3 Star Scheme view from Microsoft Analysis Server Sample Cube

Snowflake schema: The snowflake schema is a variant of the star schema model, where some dimension tables are normalized, thereby further splitting the data into additional tables. The resulting schema graph forms a shape similar to a snowflake.

The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space because a large dimension table can become enormous when the dimensional structure is included as columns. However, this saving of space is negligible in comparison to the typical magnitude of the fact table. Furthermore, the snowflake structure can reduce the effectiveness of browsing since more joins will be needed to execute a query. Consequently, the system performance may be adversely impacted. Hence, the snowflake schema is not as popular as the star schema in data warehouse design.

The main difference between the two schemas is in the definition of dimension tables. The single dimension table for item in the star schema is normalized in the snowflake schema, resulting in new item and supplier tables. The Figure 4-1 is an example of a snowflake scheme.

Fact constellation: Sophisticated applications may require multiple fact tables to share dimension tables. This kind of schema can be viewed as a collection of stars, and hence is called a galaxy schema or a fact constellation. A fact constellation schema allows dimension tables to be shared between fact tables.

4.4 Measures: Their Categorization and Computation

A data cube measure is a numerical function that can be evaluated at each point in the data cube space. A measure value is computed for a given point by aggregating the data corresponding to the respective dimension-value pairs defining the given point. Measures can be organized into three categories, based on the kind of aggregate functions used.

Distributive: An aggregate function is distributive if it can be computed in a distributed manner as follows. If the result derived by applying the function to the n aggregate values is the same as that derived by applying the function on all the data without partitioning, the function can be computed in a distributed manner. For example, count () can be computed for a data cube by first partitioning the cube into a set of subcubes, computing count () for each subcube, and then summing up the counts obtained for each subcube. Hence, count () is a distributive aggregate function. For the same reason, sum(), min(), and max() are distributive aggregate functions. A measure is distributive if it is obtained by applying a distributive aggregate function.

Algebraic: An aggregate function is algebraic if it can be computed by an algebraic function with M arguments (where M is a bounded positive integer), each of which is obtained by applying a distributive aggregate function. For example, avg() (average) can be computed by sum()/count() where both sum () and count() are distributive aggregate functions. Similarly, it can be shown that min_N(), max_N(), and standard_deviation() are algebraic aggregate functions. A measure is algebraic if it is obtained by applying an algebraic aggregate function.

Holistic: An aggregate function is holistic if there is no constant bound on the storage size needed to describe a subaggregate. That is, there does not exist an algebraic function with M arguments (where M is a constant) that characterizes the computation. Common examples of holistic functions include median(),mode() . For

instance the most frequently occurring item(s)), and rank (). A measure is holistic if it is obtained by applying a holistic aggregate function.

Most large data cube applications require efficient computation of distributive and algebraic measures. Many efficient techniques for this exist. In contrast, it can be difficult to compute holistic measures efficiently. Efficient techniques to approximate the computation of some holistic measures, however, do exist. For example, instead of computing the exact median(), there are techniques that can estimate the approximate median value for a large data set with satisfactory results. In many cases, such techniques are sufficient to overcome the difficulties of efficient computation of holistic measures.

4.5 Concept Hierarchies

A concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level, more general concepts.

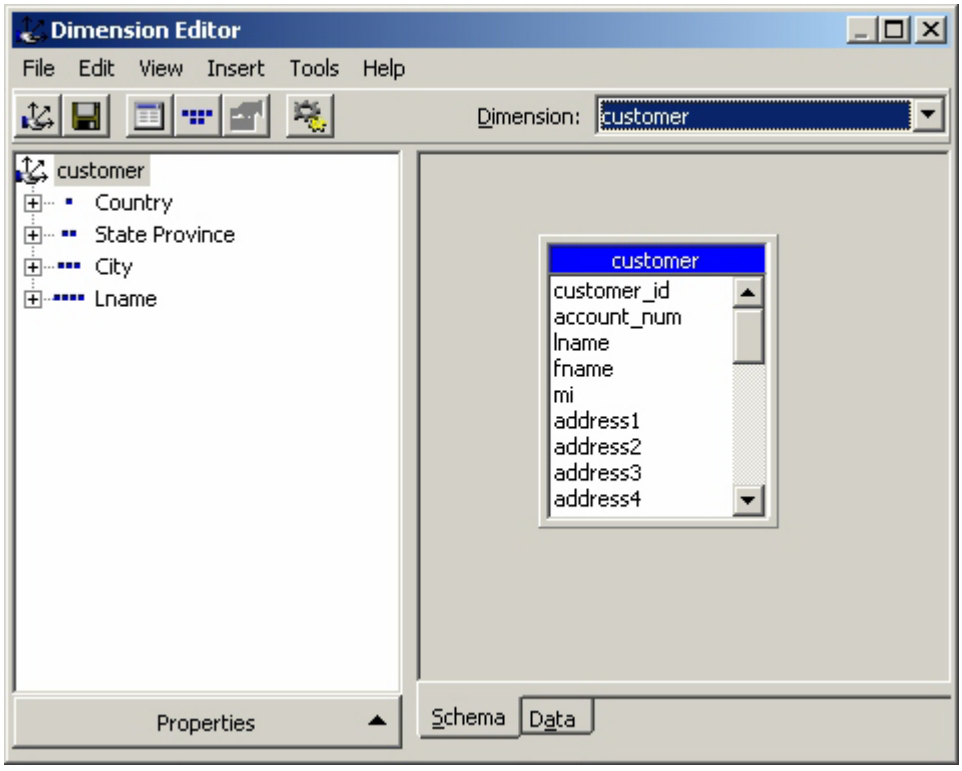


Figure 4-4 Concept Hierarchies in Microsoft Analysis Server

Concept hierarchies for customers dimension is illustrated in Figure 4-4. Many concept hierarchies are implicit within the database schema. Alternatively, the attributes of a dimension may be organized in a partial order, forming a lattice. A

concept hierarchy that is a total or partial order among attributes in a database schema is called a schema hierarchy. Concept hierarchies that are common to many applications may be predefined in the data mining system, such as the concept hierarchy for time. Data mining systems should provide users with the flexibility to tailor predefined hierarchies according to their particular needs. For example, users may like to define a fiscal year starting on April 1, or an academic year starting on September 1.

Concept hierarchies may also be defined by discretizing or grouping values for a given dimension or attribute, resulting in a set-grouping hierarchy. A total or partial order can be defined among groups of values. There may be more than one concept hierarchy for a given attribute or dimension, based on different user viewpoints. For instance, a user may prefer to organize price by defining ranges for inexpensive, moderately priced, and expensive.

Concept hierarchies may be provided manually by system users, domain experts, knowledge engineers, or automatically generated based on statistical analysis of the data distribution. Concept hierarchies can be automatically generated by selecting distinct queries from databases.

4.6 OLAP Operations in the Multidimensional Data Model

In the multidimensional model, data are organized into multiple dimensions, and each dimension contains multiple levels of abstraction defined by concept hierarchies. This organization provides users with the flexibility to view data from different perspectives. A number of OLAP data cube operations exist to materialize these different views, allowing interactive querying and analysis of the data at hand. Hence, OLAP provides a user-friendly environment for interactive data analysis.

- Roll-up: The roll-up operation (also called the *drill-up* operation by some vendors) performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.
- Drill-down: Drill-down is the reverse of roll-up. It navigates from less detailed data to more detailed data. Drill-down can be realized by either

stepping down a concept hierarchy for a dimension or introducing additional dimensions. Since a drill-down adds more detail to the given data, it can also be performed by adding new dimensions to a cube

- Slice and dice: The slice operation performs a selection on one dimension of the given cube, resulting in a subcube.
- Pivot (rotate): Pivot (also called rotate) is a visualization operation that rotates the data axes in view in order to provide an alternative presentation of the data

Other OLAP operations: Some OLAP systems offer additional drilling operations. For example, drill-across executes queries involving more than one fact table. The drill-through operation makes use of relational SQL facilities to drill through the bottom level of a data cube down to its back-end relational tables.

Other OLAP operations may include ranking the top *N* or bottom *N* items in lists, as well as computing moving averages, growth rates, and interests, internal rates of return, depreciation, currency conversions, and statistical functions. •

OLAP offers analytical modeling capabilities, including a calculation engine for deriving ratios, variance, and so on, and for computing measures across multiple dimensions. It can generate summarizations, aggregations, and hierarchies at each granularity level and at every dimension intersection. OLAP also supports functional models for forecasting, trend analysis, and statistical analysis. In this context, an OLAP engine is a powerful data analysis tool.

4.7 Starnet Query Model

The querying of multidimensional databases can be based on a starnet model. A starnet model consists of radial lines emanating from a central point, where each line represents a concept hierarchy for a dimension. Each abstraction level in the hierarchy is called a footprint. These represent the granularities available for use by OLAP operations such as drill-down and roll-up. A starnet model introducing concept hierarchies is shown in Figure 4-5

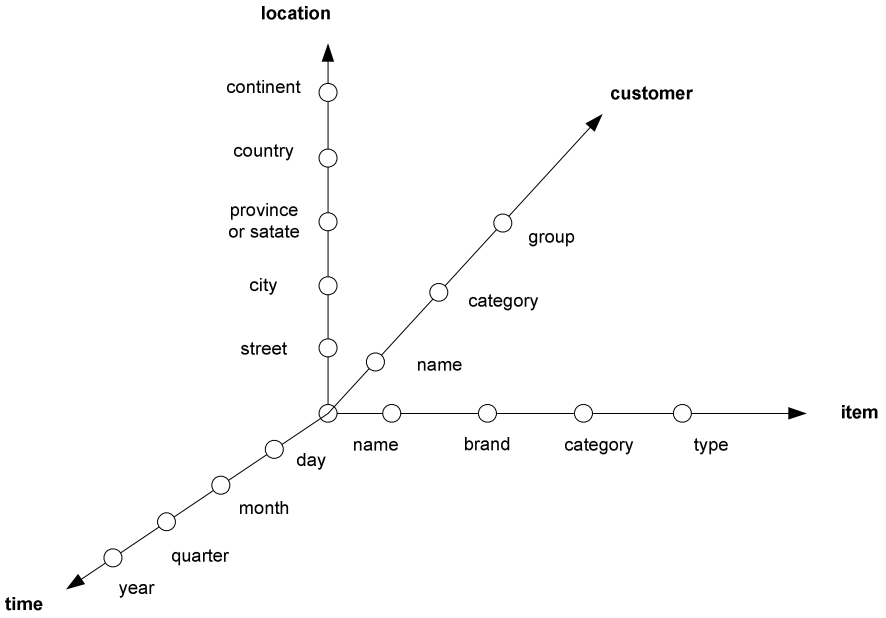


Figure 4-5 A Starnet Model Based a Customer Dimension. (Han and Kamber, 2001)

5. FUNCTIONALITIES OF DATA MINING

5.1 Characterization and Discrimination

Data Characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query. For example, to study the characteristics of software products whose sales increased by 10% in the last year, the data related to such products can be collected by executing an SQL query.

There are several methods for effective data summarization and characterization including OLAP operations.

Data Discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries. For example, the user may like to compare the general features of software products whose sales increased by 10% in the last year with those whose sales decreased by at least 30% during the same period. The methods used for data discrimination are similar to those used for data characterization.

The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help distinguish between the target and contrasting classes. Discrimination descriptions expressed in rule form are referred to as discriminant rules. The user should be able to manipulate the output for characteristic and discriminant descriptions.

5.2 Association Analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

More formally, association rules are of the form

The association rule X => Y is interpreted as "database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y"

Association analysis is usually mentioned with market basket analysis(MB) because of its wide usage in customer purchasing patterns.Market Basket (MB) Analysis refers to business-useful information that can be gleaned from aggregate associations among the different items sold in catalogs or at retail stores. The input to MB analysis is point-of-sale (POS) transactional data. The output of MB analysis is information and recommendations that exploit product associations and customer-purchase behavior. **(Han and Kamber, 2001)**

5.3 Classification

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known.**(Han and Kamber, 2001)**

The classification or supervised learning approach to data mining is very common in the business world. The human mind naturally segments things into distinctive groups. For example, people can be lumped into the classifications of babies, children, teenagers, adults, and the elderly. Classification provides a mapping from attributes to specified groupings. For example, the attribute age two years or younger can be mapped to the category babies. Once data is classified, the traits of these specific groups can be summarized.**(Groth,1999)**

The derived model maybe represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.

5.4 Prediction and Estimation

The process of prediction is straightforward. With a set of inputs, a prediction is made on a certain outcome. While the validation process uses prediction, it is really comparing known results to predictions made to calculate an accuracy level. With true prediction, the outcome to be predicted will not be known.

(Han and Kamber, 2001)

Models that are built with classification predict discrete categories. For instance, a model calculating someone's credit risk might predict them as "high," "medium," or "low." Estimation works with outcomes that have continuous values (for example, real numbers between 1 and a million). In the context of estimation, statisticians call handling of discrete value outcomes as classification and the handling of continuous value outcomes as "regression."

5.5 Cluster Analysis

Clustering is a method of grouping rows of data that share similar trends and patterns. Clustering, or segmentation, is the process of dividing a data set into distinctive groups. For supervised learning, the model takes in the independent variables, produces a guess for the dependent variable that is compared with the actual dependent value, and an error-correction is made; hence, the study is "supervised." There is no such process in clustering because there is no outcome to compare it with; hence, the study is referred to as "unsupervised." **(Groth,1999)**

For example, in the case of fraudulent claims, the records may naturally separate into two classes. One of the categories may correspond to normal claims and the other may correspond to fraudulent claims. Of course, there may be some legitimate claims that are mislabeled as fraudulent, and vice versa.

Clustering studies have no dependent variable. You are not profiling a specific trait as in classification studies. These studies are also referred to as unsupervised learning and/or segmentation.

Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known

to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

5.6 Outlier Analysis

A database may contain data objects that do not comply with the general behaviour or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are a substantial distance from any other cluster are considered outliers. Rather than using statistical or distance measures, deviation-based methods identify outliers by examining differences in the main characteristics of objects in a group.

5.7 Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behaviour changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

5.8 Visualization

Visualization is simply the graphical presentation of data. Data can sometimes be best understood by graphing it. For example, visualization techniques can easily show outliers as seen in Figure 5-1.

The process of representing data graphically is used today in most query tools. Visualization can mean much more than two-dimensional charts and maps.

An example of detecting outliers is given below, through a box plot graph outliers in sample data KDD CUP 2000 can easily be detected. The outliers in this example may refer to dirtiness in the data.

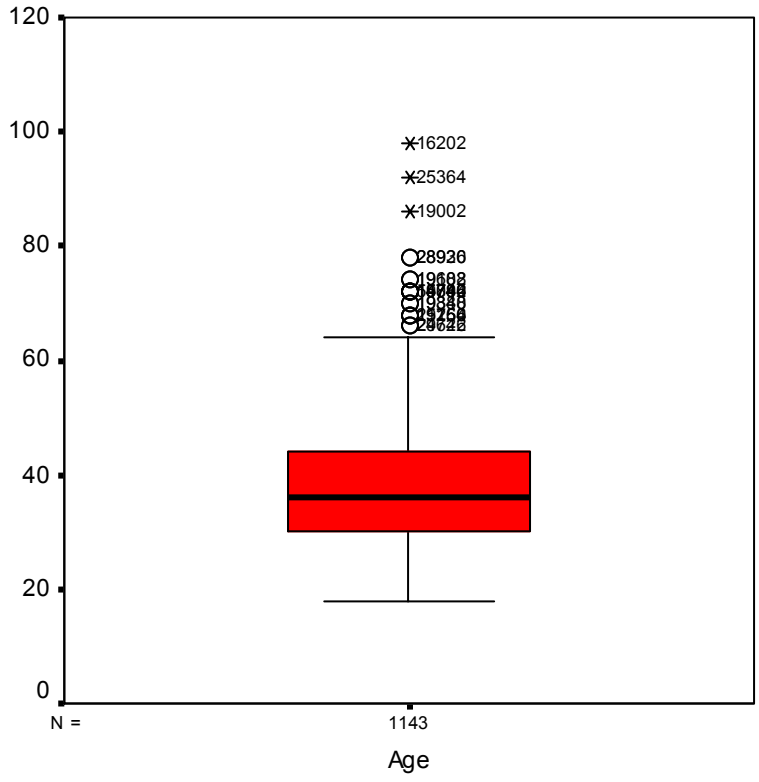


Figure 5-1 An Outlier Analysis Via Box-Plot Graph / SPSS Statistics Package

6. DATA MINING ALGORITHMS

6.1 Decision Trees

The greatest benefit to decision-tree approaches is their understandability; however, to successfully model data using the decision-tree approach, several splits may be necessary. The tree subdivides the data according to height. It may be necessary to subdivide further on the basis of age and weight to learn, for example, that short, heavier people above a certain age have a greater incidence of high blood pressure.

Below is a simplified, stepwise discussion of building a decision tree. It should be noted that there are many approaches to decision trees used today. One is a statistical approach, and CART is the best example of this approach. It uses statistical prediction in which there are exactly two branches from each nonterminal node. Another approach is where the number of branches off a nonterminal node is equal to the number of categories.

Examples of this approach are the CLS, IDS, and C4.5 algorithms. Yet another approach varies the number of nodes on a nonterminal node from two to the number of categories. This approach is exemplified by the AID, CHAID, and TREEDISC algorithms. Many vendors use a combination of these approaches. For example, Angoss Knowledge SEEKER uses a combination of algorithms.

While all decision-tree algorithms undergo a similar type of process, they employ different mathematical algorithms to determine how to group and rank the importance of different variables. For example, Quinlan, in C4.5 Programs for Machine Learning, discusses a gain-ratio algorithm that expresses the proportion of information generated by the split that is useful. The major steps in the decision tree algorithms are as follows.

- Step 1: Variables are chosen from a data source. From the variables presented in the data source, a dependent variable is chosen by the user.

- Step 2: Each variable affecting an outcome is examined. An iterative process of grouping values together is performed on the values contained within each of these variables.
- Step 3: Once the groupings have been calculated for each variable, a variable is deemed the most predictive for the dependent variable, and is used to create the leaf nodes of the tree.

Most people understand decision trees intuitively. This is the technology's greatest strength. The negative side of decision trees is the fact that they get harder to manage as the complexity of the data increases. This is because of the increasing number of branches in the tree. There is also an issue with the handling of missing data, because without a data element being present, how do you traverse a tree node dependent on that data. **(Groth, 1999)**

6.2 Genetic Algorithms

Genetic algorithms are a method of combinatorial optimization based on processes in biological evolution. The basic idea is that over time, evolution has selected the "fittest species." Applying this idea to data mining usually involves optimizing a model of the data using genetic methods to obtain "fittest" models.Genetic algorithms have often been used in conjunction with neural networks to model data.

Genetic algorithms are good at clustering data together. For example, you want to divide, or cluster, a data set into three groups. A process for doing this is discussed below.

- Step 1: For a genetic algorithm, you can start with a random grouping of data. Think of each of three clusters to be created as an organism. The genetic algorithm will have what is called a fitness function that determines if a data set is a match for one of the three "organisms" or clusters. This fitness function could be anything that identifies some data sets as "better fits" than others. As data sets are read, they can be evaluated by the fitness function to see how well they relate to the other data elements in a cluster. In our example, a fitness function could be a function to determine the level of similarity between data sets within a group.

- Step 2: Genetic algorithms have operators that allow for copying and altering of the descriptions of groups of data. These operators mimic the function found in nature where life reproduces, mates, and mutates. If a row of data in a data set is found to be a good fit by the fitness function, then it survives and is copied into a cluster. If a row of data is not a good fit, it can be crossed over to another set, or, in other words, it can be mated with other clusters to create a better fit. A cluster will alter itself to create optimized fits as new data sets are read.

Genetic algorithms solved complex problems that other technologies have a difficult time with; that having been said, however, genetic algorithms are the least understood of the approaches as well as the most "open." For example, fitness functions can vary widely. The main requirement is that a fitness function must have certain properties that allow for convergence to minimal error, yet that leaves a lot of room for varying implementations.

Genetic algorithms have often been used in conjunction with neural networks to provide a higher level of model understanding. While neural networks have often been said to be "black boxes," genetic algorithms in conjunction with neural networks can record groups of input variables that impact an outcome directly into a database, providing more detailed documentation of each neural network model. After experimenting with various models, a final model can be built by reading one of the earlier model's variable sets. **(Groth, 1999)**

6.3 Neural Networks

Neural networks are used extensively in the business world as predictive models. In particular, the financial services industry widely uses neural networks to model fraud in credit cards and monetary transactions.

Neural networks attempt to mimic a neuron in a human brain, with each link described as a processing element (PE). Neural networks learn from experience and are useful in detecting unknown relationships between a-set of input data and an outcome. Like other approaches, neural networks detect patterns in data, generalize relationships found in the data, and predict outcomes. Neural networks have been especially noted for their ability to predict complex processes.

A processing element, or PE, processes data by summarizing and transforming it using a series of mathematical functions. One PE is limited in ability, but when connected to form a system, the neurons or PEs create an intelligent model. PEs are interconnected in any number of ways and they can be retrained over several, hundreds, or thousands of iterations to more closely fit the data they are trying to model.

Processing elements, or PEs, are linked to inputs and outputs. The process of training the network involves modifying the strength, or weight, of the connections from the inputs to the output. Increases or decreases in the strength of a connection are based on its importance for producing the proper outcome. A connection's strength depends on a weight it receives during a trial-and-error process. This process uses a mathematical method for adjusting the weights, and is called a learning rule.

Training repeatedly, or iteratively, exposes a neural network to examples of historical data. PEs summarize and transform data, and the connections between PEs receive different weights. That is, a network tries various formulas for predicting the output variable for each example.

Training continues until a neural network produces outcome values that match the known outcome values within a specified accuracy level, or until it satisfies some other stopping criterion.

Each of the processing units takes many inputs and generates an output that is a nonlinear function of the weighted sum of the inputs. The weights assigned to each of the inputs are obtained during a training process (often back-propagation) in which outputs generated by the net are compared with target outputs. The answers you want the network to produce are compared with generated outputs, and the deviation between them is used as feedback to adjust the weights. **(Groth, 1999)**

The process of readjusting weights is important to increasing a model's accuracy. The number of hidden nodes can be adjusted, and, in fact, there can be multiple levels of hidden nodes, just to confuse matters. The number of inputs, hidden nodes, outputs, and the weighting algorithms for the connections between nodes determine the complexity of a neural network. In general there is a trade-off between the complexity of a neural network, its accuracy, and the time it takes to create the neural network model. Because the configuration of hidden nodes and weights is so critical

to neural networks, there are many approaches for finding the right number of hidden nodes and readjusting weights.

This is at best an introductory view of neural networks, but it does give a starting point from which to understand how they work.

The greatest strength of neural networks is their ability to accurately predict outcomes of complex problems. Neural networks are a preferred technique in performing estimation, or continuous numeric outputs, which are popular in financial markets and manufacturing. **(Groth, 1999)**

There are some downfalls to neural networks. First, they have been criticized as being useful for prediction, but not always in understanding a model. It is true that early implementations of neural networks were criticized as "black box" prediction engines; however, with the new tools on the market today, this particular criticism is debatable.

Secondly, neural networks are also susceptible to over-training. If a network with a large capacity for learning is trained using too few data examples to support that capacity, the network first sets about learning the general trends of the data. This is desirable, but then the network continues to learn very specific features of the training data, which is usually undesirable. Such networks are said to have memorized their training data, and lack the ability to generalize. Commercial-grade neural networks today have effectively eliminated overtraining through bootstrapping holdout (test) samples, and by monitoring test versus training errors.

Over-training can be measured by periodically checking the results of your test data set. Early stages of a training session yield lower error measurements on both the training and the test data. This continues unless the network capacity is larger than need be, or unless there are too few data sets in the training file. If at some point during learning, your test data begin to produce worse results, even though the training data continue to produce improved results, over-training is occurring.

Another issue with neural networks is training speed. Neural networks require many passes to build. This means that creating the most accurate models can be very time consuming. It is only fair to mention that all regression techniques require time to converge; and, while back propagation is slow, training neural networks can be sped up dramatically with methods like conjugate gradient.

6.4 Statistics

Strengths of statistical approaches are that not only are these approaches accurate, they are well understood and widely used. Statistical approaches are viewed by many to be the "truest" form of data mining, and in fact, many data mining techniques make use of statistical techniques that have been around for many years. CHAID, a popular decision tree approach, uses the Chi Square metric. Association algorithm use the statistical metrics of support and confidence, and clustering techniques use statistical metrics like the K-Means algorithm. Bayesian Networks use the Bayes Theorem of Probability.

The biggest criticism of statistics has been the perceived difficulty of using it effectively. many business professionals are confused by the terminology used in the statistic. **(Groth, 1999)**

7. INDUSTRY APPLICATIONS OF DATA MINING

7.1 Data-Mining Applications in Banking and Finance

Data mining has been used extensively in the banking and financial markets. In the banking industry, data mining is heavily used to model and predict credit fraud, to evaluate risk, to perform trend analysis, and to analyze profitability, as well as to help with direct marketing campaigns.

In the financial markets, neural networks have been used in stock-price forecasting, in option trading, in bond rating, in portfolio management, in commodity price prediction, in mergers and acquisitions, as well as in forecasting financial disasters.

(Dahlan and others, 2002)

- Identify patterns of fraud
- Identify loyal customers
- Predict customers likely to change credit card companies
- Find hidden correlations between different financial indicators
- Identify stock trading rules from historical market data

7.2 Data-Mining Applications in Retail

Slim margins have pushed retailers into embracing data warehousing earlier than other industries. Retailers have seen improved decision-support processes lead directly to improved efficiency in inventory management and financial forecasting. The early adoption of data warehousing by retailers has given them a better opportunity to take advantage of data mining. Large retail chains and grocery stores store vast amounts of point-of-sale data that is information rich. In the forefront of the applications that have been adopted in retail are direct marketing applications.

- Reorganizing such that the products that sell together are placed together to induce impulse buying.
- Deciding promotion schemes
- Optimum Inventory decisions
- Find buying patterns
- Learn associations among customer demographics
- Predict customer responses to advertising
- Perform Market Basket analysis

And also in Internet marketing;

- Targeting of advertisements
- Personalization of web pages
- Association of items likely to be viewed or purchased together
- Classification of articles automatically
- Clustering/segmentation of groups sharing common characteristics
- Estimation of missing data
- Prediction of future behavior

(Groth, 1999)

7.3 Data-Mining Applications in Healthcare

Data mining has been used extensively in the medical industry already. For example, Neuro Medical Systems used neural networks to perform a pap smear diagnostic aid. Vysis uses neural networks to perform protein analysis for drug development. The University of Rochester Cancer Center and the Oxford Transplant Center use KnowledgeSEEKER, a decision tree technology, to help with their research. The Southern California Spinal Disorders Hospital uses Information Discovery to data mine. Information Discovery quotes one doctor as saying: "Today alone, I came up with a diagnosis for a patient who did not even have to go through a physical exam." **(Groth, 1999)**

7.4 Data-Mining Applications in Telecommunications

In recent years, the telecommunications industry has undergone one of the most dramatic makeovers of any industry. The U.S. Telecommunications Act of 1996 allowed Regional Bell Operating Companies (RBOCS) to enter the long-distance market and offer "cable-like" services. The European Liberalization of Telecommunications Services, effective January 1, 1998, liberalized telecommunications services in Europe, and offers full competition among participating European nations. Sixty-eight nations liberalized their telecommunications market on January 1, 1998 to coincide with the European commitment based on the World Trade Organization's Telecommunications Agreement.

Not only has there been massive deregulation, but in the United States, there has been a sell-off by the FCC of airwaves to companies pioneering new ways to communicate. The cellular industry is rapidly taking on a life of its own.

With the hyper-competitive nature of this industry, a need to understand customers, to keep them, and to model effective ways to market new products to these customers is driving a demand for data mining in telecommunications where no demand existed in distant memory.

Several companies offer products to combat customer churn. For example, RightPoint Corporation focuses on data-mining issues in the telecommunications industry and, in particular, customer retention or churn. Industry experts have pointed out that the cellular telephone market experiences a 30% churn rate in the United States. A report by Digital Equipment Corporation, produced by Evan Davies and Hossein Pakraven in September 1995, quantifies the cost of customer churn. In their report, they estimate that the cost of acquiring new customers is as high as $400 for each new subscriber. Data visualization is another area with many strategic uses in telecommunications. **(Groth, 1999)**

7.5 Data Mining Applications in Manufacturing

In the competitive environment, manufacturers can no longer rely on low prices, high quality and on-time delivery alone to keep them on top. These attributes were advantages a decade ago but now they are just requirements to stay in the business. And the rules of the business are constantly changing. Manufacturers face increasing globalization, more competition than ever, and customers whose demands reflect their own knowledge and expectations of a global market. In order to take competitive advantage they have to deploy knowledge management systems in their manufacturing systems. Data Mining is one of the core technologies in the value chain. It has many application areas in Manufacturing including;

- Demand Planning
- Quality Improvement
- Supplier Relationship Mgmt.
- Supply Chain Analysis
- Value Chain Analysis
- Warranty Analysis

More specifically;

- Selection of Materials and Manufacturing Processes
- Time Series Analysis and Data Mining
- Fault Diagnosis
- Data Mining for Preventive Machine Maintenance
- Manufacturing Knowledge Acquisition with Data Mining
- Process and Quality Control
- Predicting Assembling Errors
- Process Analysis
- Operational Manufacturing Control (For example: schedules that learn, the effect of local dynamic behavior on global outcomes)

- Dynamic indexing and retrieval of manufacturing information in knowledge bases summarization and abstraction of large and high-dimensional data (Self Organizing Map (SOM) based data visualization methods)
- Adaptive Human-Machine Interface for Machine Operation
- Feature Selection and Dimensionality Reduction of Manufacturing Data
- Extracting Process Yield Classes with SOM
- Feature Recognition with SOM and Neural Networks
- Cutting tool-state classification for tool condition monitoring
- Data Mining for Capturing Best Manufacturing Practice
- Learning in the context of Robotics (For example: navigation and exploration, mapping, extracting knowledge from numerical and graphical sensor data)
- VLSI implementation of Neural Networks and Fuzzy Systems

(Groth, 1999)

8. DATA MINING ON THE INTERNET

8.1 World Wide Web and Web Datawarehouses

The World Wide Web (known as “WWW” or “Web”) is growing at a phenomenal rate. The reason for the Web’s success is largely due to its simplicity. It allows users to publish and retrieve information easily via the hypertext interface. Another important feature is its compatibility with other existing protocols, such as gopher, ftp, net news, telnet, etc. Moreover, it provides users with the ability to browse multimedia documents in an open environment available on many different platforms, requiring little cooperation between information providers and users.

(Hongjun and Feng, 1998)

The Web revolution has certainly not replaced the need for the data warehouse. In fact, the Web revolution has raised everyone's expectations much higher that all sorts of information will be seamlessly published through Web browser interfaces. The audience for data warehouse data has grown from internal management to encompass customers, partners, and a much larger pool of internal employees. The Web's focus on the "customer experience" has made many organizations much more aware of learning about the customer and giving the customer useful information.

The Web revolution has propelled the data warehouse out onto the main stage, because in many situations the data warehouse must be the engine that controls or analyzes the Web experience. In order to step up to this heightened responsibility, the data warehouse must adjust. The nature of the data warehouse needs to be somewhat different than it has been for the past decade. Figure 8-1 show the relationship of the customer to the Website and to the Webhouse. **(Kimball and Merz, 2000)**

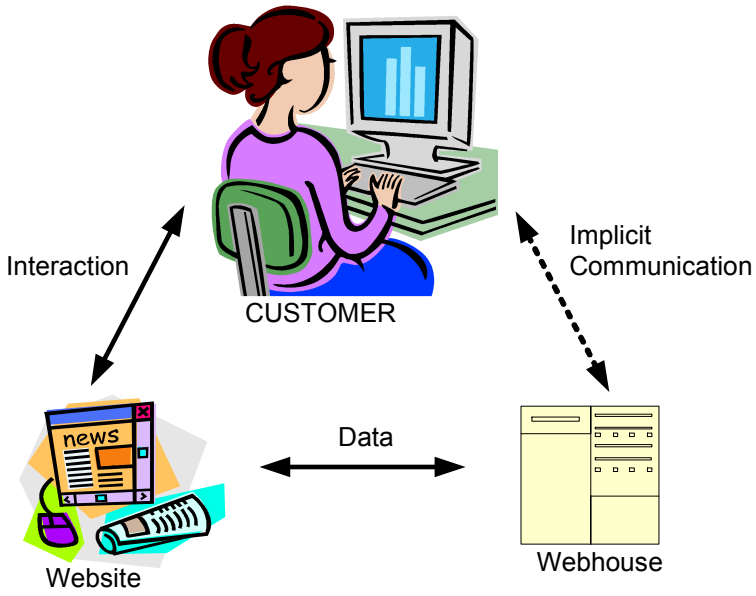


Figure 8-1 The Customer, the Website, and the Webhouse (Kimball and Merz, 2000)

8.2 Data Mining on the Internet

Data mining on the Internet, commonly called web mining, needs to take advantage of the content of documents, but also of the usage of such resources available and the relationships between these resources. Web mining, the intersection between data mining and the World-Wide Web, is growing to include many technologies conventionally found in artificial intelligence, information retrieval, or other fields. Agent-based technology, concept-based information retrieval, information retrieval using case-based reasoning, and document ranking using hyperlink features and usage (like CLEVER) are often categorized under web mining. Web mining is not yet clearly defined and many topics will continue to fall into its realm.

"Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World-Wide Web ".

(Zaiane, 1999)

Figure 8-2 shows a classification of domains Mining in the World-Wide Web field, there are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their

descriptions. Web document text mining, resource discovery based on concepts indexing or agent-based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the World-Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs. **(Zaiane, 1999)**

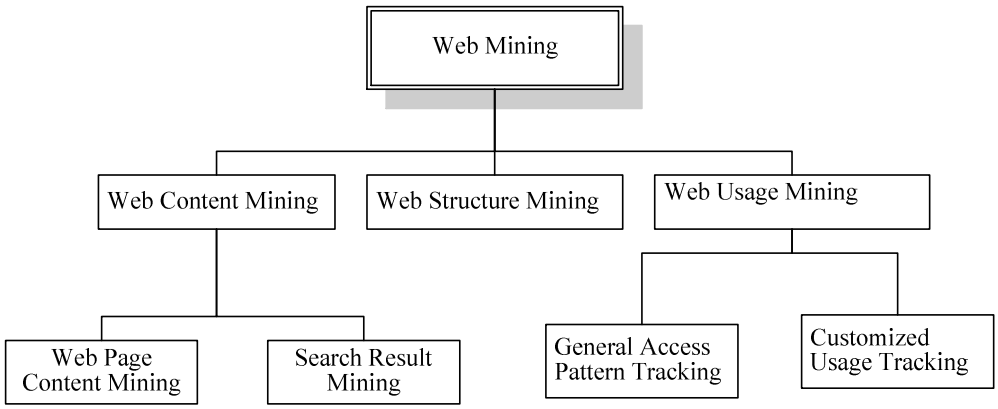


Figure 8-2 Taxonomy of Web Mining Techniques. (Zaiane, 1999)

8.2.1 Web Content Mining

Most of the knowledge in the World-Wide Web is buried inside documents. Current technology barely scratches the surface of this knowledge by extracting keywords from web pages. This has resulted in the dissatisfaction of users regarding search engines and even the emergence of human assisted searches on the Internet. Web content mining is an automatic process that goes beyond keyword extraction

8.2.2 Web Structure Mining

The World-Wide Web can reveal more information than just the information contained in documents. For example, links pointing to a document indicate the popularity of the document, while links coming out of a document indicate the richness or perhaps the variety of topics covered in the document. This can be compared to bibliographical citations. When a paper is cited often, it ought to be important. The PageRank and CLEVER methods take advantage of this information conveyed by the links to find pertinent web pages. **(Zaiane, 1999)**

Also web structure mining can be applied to determine “Web Site Effectiveness”. Web administrators can rely on structure mining results when designing the layout of

a Web site. E-retailers can develop the look and feel of the Web site and personalize online content. **(Doherty, 2000)**

8.2.3 Web Usage Mining

Despite the anarchy in which the World-Wide Web is growing as an entity, locally on each server providing the resources there is a simple and well structured collection of records: the web access log. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the design of this colossal collection of resources. There are two main tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking. The general access pattern tracking analyzes the web logs to understand access patterns and trends. These analysis can shed light on better structure and grouping of resource providers. Many web analysis tools exist but they are limited and usually unsatisfactory. Applying data mining techniques on access logs unveils interesting access patterns that can be used to restructure sites in a more efficient grouping, pinpoint effective advertising locations, and target specific users for specific selling ads. Customized usage tracking analyzes individual trends. Its purpose is to customize web sites to users. The information displayed the depth of the site structure and the format of the resources can all be dynamically customized for each user over time based on their access patterns. One innovative study has proposed such adaptive sites: web sites that improve themselves by learning from user access patterns. While it is encouraging and exciting to see the various potential applications of web log file analysis, it is important to know that the success of such applications depends on what and how much valid and reliable knowledge one can discover from the large raw log data. Current web servers store limited information about the accesses. Some scripts custom tailored for some sites may store additional information. However, for an effective web usage mining, an important cleaning and data transformation step before analysis may be needed. **(Zaiane, 1999)**

8.3 Decisions Based Upon Clickstream Analysis

8.3.1 Customizing Marketing Activities by Identifying Customers

The goal of identifying and recognizing the customer is to establish a more meaningful relationship than is possible with an anonymous transaction. We can begin to personalize our interaction with the customer when we see that the customer has returned to our site. As our experience with the identified customer grows, we will offer different deals, different opportunities, and different "customer faces" to:

- High-profit customers vs. low-profit customers,
- New customers vs. returning customers, and
- Reliable product keepers vs. frequent product returners.

Identifying the customer entering our Website is the most basic single requirement of the clickstream analysis. Identifying the customer is the foundation for almost all of the decisions. There are four levels at which the customer can be recognized according to Kimball and Merz as they introduced in their book. the customer, in order of increased knowledge of who the customer really is. (**Kimball and Merz, 2000)**

1. A persistent identifier that only tells us that a Web browser on a particular computer is engaging in a session at this moment. This level-1 anonymous tag cannot be used as a reliable guide to identifying a future session from the same computer.

2. A persistent identifier that tells us the same Web browser on a particular computer has returned for a repeat session.

3. A persistent identifier that tells us a particular human being has returned to our Website.

4. A permanent and specific personal identifier that tells us reliably that a known customer has returned to our Website. In this case we know the true name of the customer and some of the customer's demographic information. We may have obtained the demographic information in a cooperative way by asking the customer, or we may have purchased demographic data from a data supplier by linking through the customer's known name and address.

Even level 1 is useful, if that is all we can get. Level 1 at least allows us to track an anonymous session and perhaps to classify the session as successful or unsuccessful. Levels 2 and 3 are significantly better because we can measure return visits. A return visit is very significant because it means the customer is interested in our site. We are providing something useful or interesting. The customer has made our site part of his or her life.

Level 4 is the most strategic level. When we have some idea who we are dealing with, we can be far more responsive and present far more of a customized interface. In many cases, we have a level-4 understanding of who the customer is because the customer has actually made a purchase or has used our services in a way that reveals their identity.

8.3.2 Targeting Marketing Activities by Clustering Your Customers

Once we have at least a level-2 identification of the customer, we can measure certain characteristics of the return visits to our Website. A classic, simple way of clustering customer behavior is to accumulate three basic measures: recency, frequency, and intensity. (**Kimball and Merz 2000)**

- Recency is how many days it has been since we last saw the customer on our Website.
- Frequency is how many times we have ever seen this customer on our Website,
- Intensity is the grand total of the customer's purchases or some other quantitative measure of our basic Website objective. Sometimes intensity is called "monetary."

Driving marketing activities needs clickstream and sales transaction data marts. Clustering customers by recency, frequency, and intensity is attractive because we may be able to use the clickstream data by itself to perform the whole analysis. Certainly, the recency and frequency measures can be measured in the clickstream. The intensity measure may or may not be available directly in the clickstream. If the intensity measure is directly related to a page event in the Web server log, then it is in the clickstream. But if the intensity measure is the total volume of purchases as recorded in a companion transaction system fed, but not captured, by the Web server,

then we will have to generate the intensity measure by drilling across to the sales transaction data mart. Similarly, we might decide that the intensity measure such as total sales includes sales that didn't happen exclusively through the Web interface. In this case we are absolutely committed to drilling across to the non-Web source to complete the clustering measures.

Cluster analysis can be much more sophisticated than simply accumulating recency, frequency, and intensity. If we have a good verbose demographic description of the customer with many textual and numeric attributes, then we have a full-fledged data mining problem. If we have thirty demographic descriptors of the customer, we do not build a thirty-dimension cube and look at it graphically. Fortunately, there are lots of powerful data mining tools that can sort through a large number of demographic descriptors and numerical measures and advice the analyst which of these variables combine to show interesting clusters. Clustering and data mining techniques can be used to directly recommend marketing decisions. Rather than simply clustering customers relative to revenue or profit, customers can be clustered according to their history, and hence their likelihood, of responding to certain kinds of promotions. We use these techniques to decide how to cross-sell, up-sell, and create promotions for each specific customer.

8.3.3 Evaluating Cross-Link Reference

Every arriving page request from the Web will normally identify the referring site from which the page request was launched. We assign the newly arrived prospective customers to known clusters of customers whose behavior we understand. The customers coming to us from the link in question perhaps can be rated by likelihood to buy, by revenue, by profit, by lifetime value, by propensity to return the product, or by propensity to invoke costly support. All of these factors would allow us to make an informed decision whether to encourage or support a referring cross link.

In many cases, of course, we are paying the referrer for each arriving Website hit. Paid Website referrals can be abused. Without careful monitoring, we can't be sure just what the context of the referral might have been. We do know we have to pay the referrer when they dump a "customer" in our lap. So let's try to develop an informed opinion about the worth of these new customers.

Deciding whether to encourage or support a referring cross-link needs the clickstream data mart.

Usually if the page request arriving at our site has come from a search engine, the search string used by that engine is also available. This scenario is even more compelling if we have our own intra-site search facility because then we will know much more about the context of the search. In either case, if we can parse the search string, then the context of the search may be understandable. We may be able to add meta tags and verbiage to parts of our Website to draw in more productive hits, and we may be able to remove items that are causing nonproductive hits.

8.3.4 Determining Whether a Customer Is About to Leave

There are several situations where we would like to know if a customer is about to leave us. During an actual session, we may be able to tell that someone can't find what they want, and that they may be frustrated. If we detect this in real time, we may be able to present a custom Web page and ask them directly what is wrong. The diagnosis of a frustrated on-line customer can be based on how they arrived at our site, especially if the search criteria from a search engine are available. The speed and breadth of the customer's page requests may be another good indicator. If the customer is clicking very quickly, they aren't reading the page. Of course it helps to know if this is a repeat customer. A repeat customer may know exactly where to click to .get to a destination. But a new customer is probably jumping from place to place because they aren't finding anything useful. The decision that we are trying to make is whether to intervene.

Deciding whether a customer is about to leave us requires the clickstream, sales transaction, and customer communication data marts.

A more serious form of abandonment is the established customer whose trust or whose interest we are about to lose. In this case, a more complex pattern must be analyzed. We might look in the clickstream for recency, as well as frequency and intensity in the most recent time periods. We might also look for unsuccessful visits, where the established customer came to our site but left without completing a transaction. (**Kimball and Merz 2000)**

8.3.5 Determining Whether a Particular Web Ad Is Working

We would like to measure whether an ad on our Website or on a remote Website leads to increased sales, increased profits, and better customers. In this case, we assume the customer does not explicitly interact with the ad by checking on it or treating it as a link. Explicit interactions are easy to measure and make decisions about. The more subtle problem we are trying to measure here is a soft causal effect. This is similar to measuring advertising effectiveness in conventional media such as radio, television, or newspaper. In these cases, we don't know whether our ad registered on the consciousness of the customer or whether the ad led to the customer seeking our products.

In the world of conventional media, ad campaigns can often be measured only by an indirect increase in sales that "seems" to be in response to the ad. In a complex marketplace where we and our competitors are bombarding the customers with overlapping and conflicting stimuli, it is often a guess or an article of faith that an ad campaign has done very much. The most direct measures of ad effectiveness are surveys of brand awareness created by the ads. But it is still a leap of faith that brand awareness is the reason for increased sales. (**Kimball and Merz, 2000)**

8.3.6 Determining If Custom Greetings Are Working

The generation of custom greetings is a significant decision in an e-commerce environment because it requires a lot of infrastructure to do well. A custom greeting may be a completely precalculated marketing message or there may be a simple cache of summary information that can generate a few predictable messages about the customer's account, the customer's last order, any backorders, and other opportunities, or special deals we think would appeal to the customer. A very important kind of custom greeting is a cross-selling or up-selling proposal. Cross-selling is selling a product or service belonging to a family of comparable products. For instance, a bank may propose that a customer open a savings account to accompany a checking account. Up-selling is selling a product or service of significantly more value than the ones already used. The bank may propose a home mortgage or a small business loan to customers with the right profile and history.

In any case, the cache of custom greetings needs to be updated frequently because it needs to reflect the most current reality. The main business transaction server needs to update the cache whenever a meaningful transaction takes place. The customer may want to see the status of a transaction seconds after it was posted. Also, the main relational database stores of historical data will periodically create custom greeting such as the cross-sell and up-sell proposals based on large groups of customers and based on more significant time histories than are available on the hot response cache server.

The decision whether custom greetings are working is similar to the decision whether ads are working. If the custom greeting is interactive, then we can directly measure a kind of impulsive response. But for the noninteractive greeting and for the delayed response, we have the same soft causal issues described in the previous section. We look at sales to those people exposed to the greeting, and we look at timing of such sales relative to the greeting.

Deciding if custom greetings are working needs the clickstream and sales transaction data marts. Cross-selling and up-selling needs a "core" revenue data mart spanning multiple lines of business.

Custom greetings are more powerful than many other forms of ads because we control who sees the greeting. We can easily create control groups of those customers who see a greeting and those who don't. We can be much more confident that small differences in behavior are due to the greeting or the lack of the greeting.

8.3.7 Determining If a Promotion Is Profitable

A deeper issue in marketing is to decide whether a "promotion is profitable." In this case, a promotion is an entire marketing campaign, including development costs, media costs, and all the financial incentives, including temporary price reductions that we pass on to the customer as part of the promotion. It is also very important to realize that when the boss walks into the marketing department and asks whether the "promotion is profitable," the boss isn't really asking whether the incremental transactions recorded as part of the promotion were individually profitable, but several much harder questions:

• Was running the promotion better than not running the promotion? In other words, was the overall profit of the company higher as a result of running the promotion than if we had not run the promotion? Answering this question requires guessing a "baseline" level of sales that would have taken place if only we had not run the promotion.

- Did the promotion cannibalize other products that we sell? In other words, did we simply transfer sales from regularly priced products to temporarily low-priced products?
- Did regular sales drop noticeably either before or after the promotion?
- Did we increase the size of our market, even if we did not show an obvious increase in profit?

Deciding if a promotion is profitable needs the clickstream, sales transaction, promotions management, and competitive intelligence data marts.

8.3.8 Responding to a Customer's Life Change

In many businesses, major changes in the customer's life create the opportunity to deepen the relationship with the customer and to extend the range of products and services sold to that customer. Significant life changes include

- Marriage or divorce
- Having a child
- Going to college; sending a child to college
- Buying or selling a house
- Moving to a new city
- Becoming a care taker for an elderly relative
- Retirement
- Major health changes

8.3.9 Determining the Profitability of Web Business

At some point, every Web business needs to step back and ask the basic profit question. If the entire enterprise is a Web business, then answering this question is

easier than if only a small part of enterprise is a Web business. The company that is totally committed to the Web doesn't have to apportion revenues and costs between Web and non-Web activities. In this case, the annual report itself will show whether the Web business is profitable. However, in all businesses that are trying to see if their Web activities are profitable, one must break the analysis of profitability down to a very low granularity so that many different views of profit can be constructed. Since so many Web-enabled businesses have a strong customer focus, it is very desirable to ask:

- Which groups of customers are profitable?
- Which groups of products or services are profitable?
- In which time periods are we profitable?
- Which promotions are profitable?
- Are we profitable on an incremental basis? On a fully burdened basis
- And finally, is the Web business profitable overall?

The secret to answering all these variations of the profit question to build a complete activity-based profit and loss statement at an extremely low, granular level of the business. For a Web-enabled business we recommend building this P&L statement at the grain of the individual customer session. We will allocate as much of the costs of the business as we can down to this very low level. In some cases, this allocation process will be painful or controversial, but our efforts will repaid by being able to answer all of the questions in the preceding list. one of the cost components. This dilemma is reflected in the following tip. (**Kimball and Merz 2000)**

9. APPLICATION: CLICKSTREAM DATA ANALYSIS ON A WEB RETAILERS DATA

9.1 Problem Definition

As the internet matures as a medium for doing business, companies are trying to move beyond the basics of business transactions into a deeper level of understanding of what is occurring at their web site. Primary to this is the understanding of how their customers are interacting with the site, which includes not only navigation patterns, but other customer information such as demographic data and buying habits. (**Brainerd and Berry**, **2001)**

A key component here is not only seeing the navigation patterns of the customer, but combining that with other information about that customer, including demographic and purchase information. It is this deeper level of understanding of the customer and how the customer is behaving in the store that allows the retail manager to make more informed business decisions. Some of the strategic decisions were summarized in the previous section. In the virtual marketplace, there is no direct, physical interaction with customers, so we must come up with new ways of understanding customers' behavior in a web site.

The assessment of user profiles is an issue of major interest for web applications. User modeling is being investigated for a long time in different contexts and for a variety of domains, ranging from intelligent tutoring systems to recommendation agents in e-commerce. The preferences of and further information on web site visitors can be obtained either by requesting input from the users or by drawing conclusions based on the users observed behavior. **(Masand, 2000)**

To perform a real case study, KDDCUP2000[3] data was downloaded and handled. The data was collected with a dedicated software Blue Martini which run at the application server of a web retailer focused on legwear and legcare products. Personal information within the data was removed by the organization committee of the competition in order to prepare the data for the competition. The details about the competition can be found in Appendix B.

Data consists of two separate training sets including click stream sessions and order sessions. Each row in clicks data represents one page view of the user. Each row in orders data represents an order line which can be a part of an order. Using OLAP and data mining functionalities, interesting patterns that may serve the decision maker were investigated. The competition question "Given a set of purchases over a period of time, characterize visitors who spend more than $12(order amount) on average order at the site. was the main focus in order to predict valuable customers. Although we tried to answer the question above we made the analysis independent from the competition guidelines.

9.1.1 Data Structure

The data consists of approximately 4,000 customers who produced 900,000 requests (clicks), 400,000 sessions, and 2,000 orders on 1,000 distinct products over a two month period between 29/1/2000 and 29/3/2000. Demographic data, such as gender, was collected via an online survey form that was filled out by site users during a registration process, and supplemented with data from Acxiom, a third-party data supplier. The competition version of data can be downloaded from http://www.ecn.purdue.edu/KDDCUP/ .

The data was collected using Blue Martini Architecture (infrastructure that captures web logs). The data dictionary, which helped us to identify the variables, .can be seen in Appendix B

[3] A competition within KDD-2000 (The Sixth ACM SIGKDD International Conference on Knowledge Discovery and DataMining), Boston, MA, August 2000. The data is available for download from www.bluemartini.com/kddcup2000.

9.1.2 Importing Data to Database Server

This process took much more time and effort than we had expected in the overall mining process. The data was in plane text format. The first data file (1,148.6 MB) held the clickstream sessions in which the user explored the site. The second data file (4.9MB) held the order sessions in which the user placed his/her order.

The problem was the amount of data. Data could not be explored by office automation solutions like spread sheets, word processors. A more effective and powerful environment was needed to explore and prepare the data. Microsoft SQL Server, which is a database server, was used to import and maintain data. The DTS Data Import Wizard of Microsoft SQL Server was used to import the comma delaminated text data. There had been many problems during the import process because the import wizard had problems with last service pack. (SQL Server Service Pack 3). To fix this problem the service pack was set up after the data was imported. All of the data was imported as variable string format. (varchar 255).

The next step was to combine the data and data column names which were in separate text files. The stored procedure command set was used to rename the column names. The command format was prepared in a spreadsheet solution using the appropriate string concatenating formula. A segment of the formulas and commands applied can be examined in Appendix B.

9.2 Data Exploration and Understanding

The data dictionary given by the competition organization team was explored to clarify variables.(Appendix B). We noticed while exploring the data that the customer specific variables in clicks data were empty. Orders data file had customer information specific columns with huge amounts of missing values. The tables were not normalized and there were more than one variable referring the same content. The frequently asked questions section of the KDDCUP2000 contest was explored to understand the structure of the data. The problems were;

- Some of the data variables are only used in certain periods in time. This is caused by the structural changes in the data collector questions.
- There were outlier cases in clicks data caused by robots and web crawlers.

- Boolean (True /False) variables were suspicious to be trusted because they had no null values. They had been possibly set to have a default value which may mislead the analyzer. These notifications were taken into account during data preparation and analysis processes.

9.3 Data Preparation

9.3.1 Data selection

In order to solve our business problems including customer profiling and customer segmentation, we needed customer demographic information. The order data set had a variety of demographic information about the customers. The site has purchased demographic data of its customers from a marketing research company (Acxiom Data). The clicks data had no customer demographic information despite the click sessions which resulted with an order could be matched from orders data file.

The data selection process was applied to exclude the variables that were not relevant to our business decisions. Some important variables also had to be dropped because of the missing values. The missing values were examined through the bar graphics seen in Figure 9-1. Bars in the graph represent the percentage of the valid cases in the data for the specific variable. The same analysis could not be done for clicks data by a spreadsheet solution because of the volume of the data.

To solve the problem caused by structural changes in questionnaire questions, the data was grouped by the variable “lastupdate_date”. “Last update date” variable referred to the date of the structural changes made in the questions. The case 2000-05-04 held nearly 98% of the distinct cases. The rest of the data was removed from the database since it weakened the data quality.

9.4 Data Cleaning

Some of the variables had invalid characters and symbols. To set up a valid structure and format for our data environment these characters had to be removed. This was done by using transactional SQL commands. Samples to these commands were given in Appendix C. The outliers from the clicks data were eliminated by transactional SQL commands.

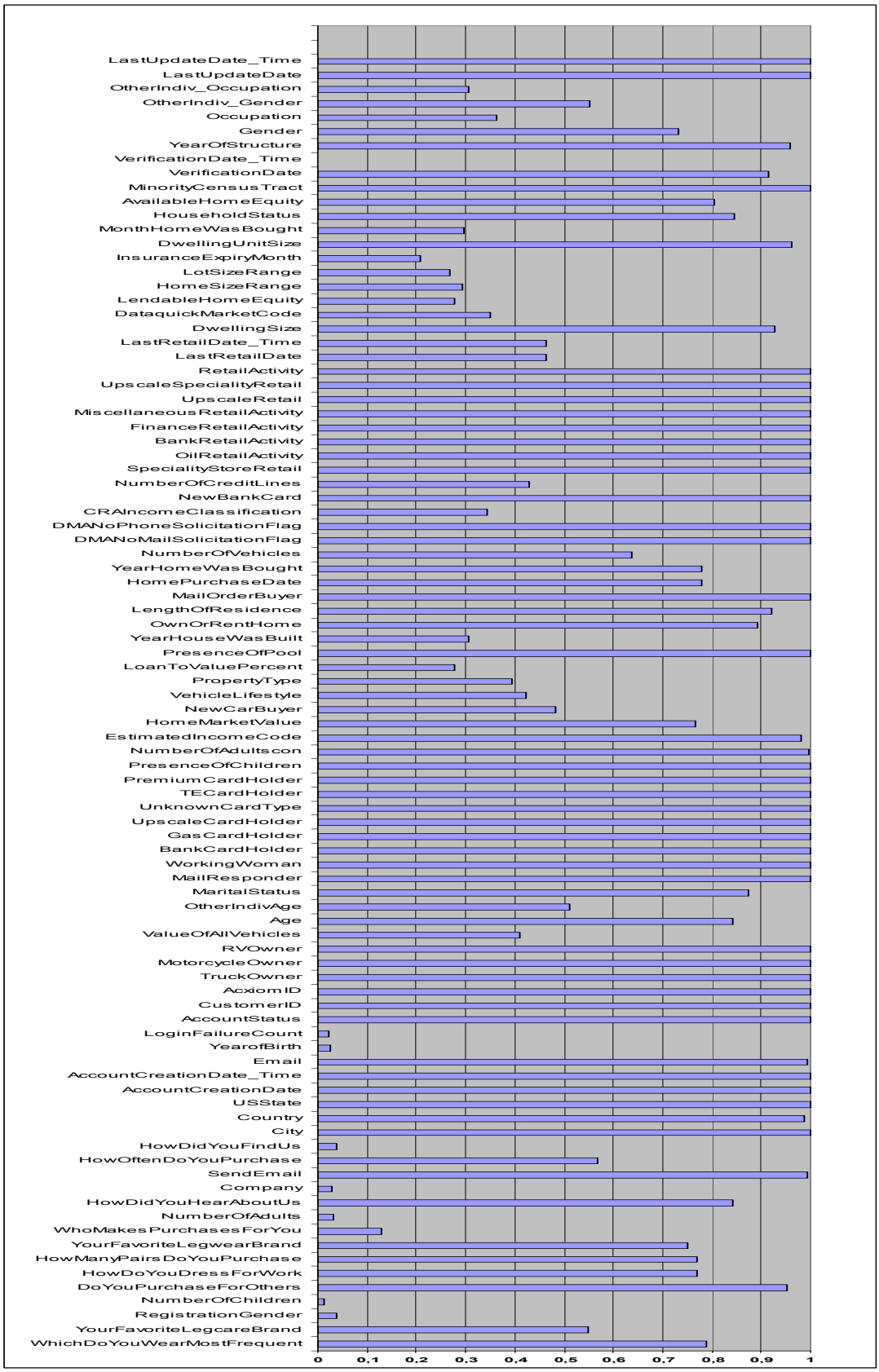


Figure 9-1 Analysis of Missing Values using Microsoft Excel.

9.5 Data construction and Integration

The data was transformed into a relational structure in SQL Server. The base table constructing the relational structure was the orders data. Customer table, Orders fact table, Product table were created from customers data. To benefit from the OLAP technology tables were designed to fit OLAP cube structure. The tables were designed as seen in Figure 9-2.

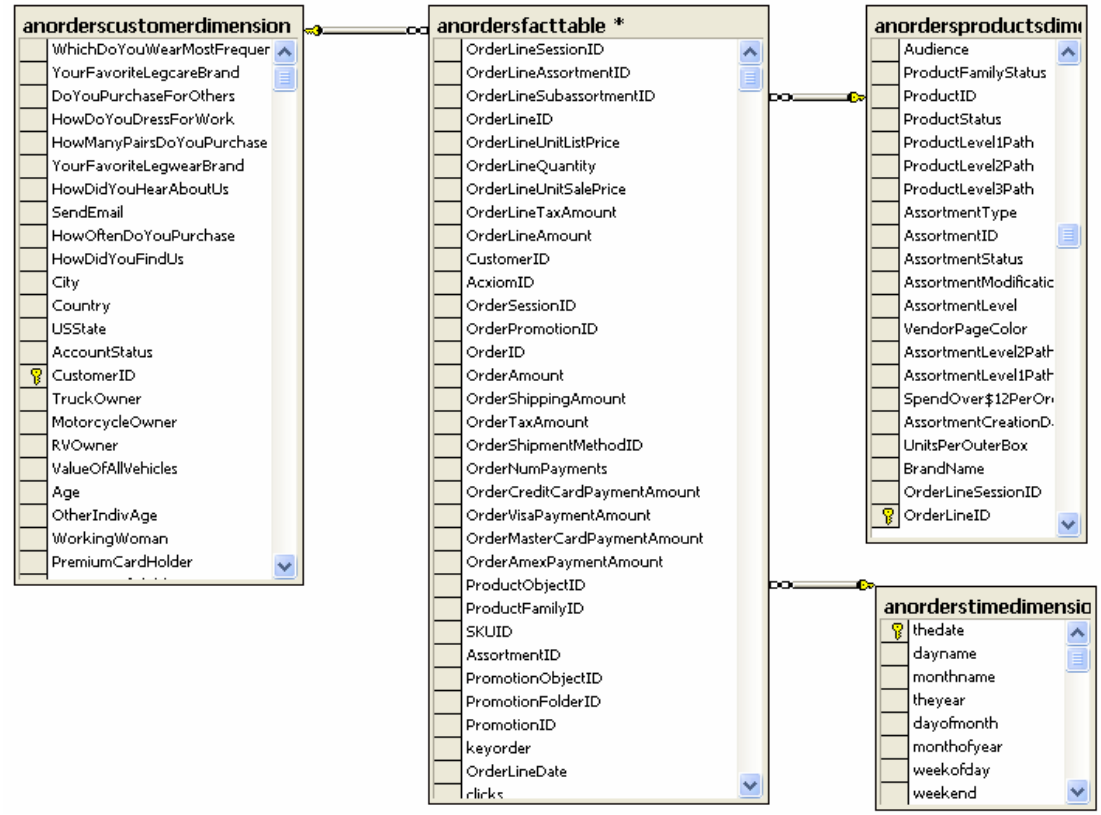


Figure 9-2 Relational Table Diagram in SQL server

9.5.1 Customer Table

Customer table was created from orders database with the orders data. Repeated records were eliminated. CustomerID field was set to be primary key (unique identifier). The table was designed to be a dimension for OLAP analysis. Variables were selected to investigate customer preferences and customer characteristics. Numeric fields such as “age”, “number of cars owned” were transformed to numeric variables using SQL Server table designer.

9.5.2 Products Table

Products table was also created from orders database. This table contained the product related fields in the order placed by customer. Eliminating repeated records was not as easy as it had been in the customer table because the variable "productID" had significant number of missing values. The repeated records were left as they were in the table since many of the OLAP tools provided distinct data analysis. The OLAP tool we had used, Microsoft Analysis Server had the ability to select distinct cases so this situation had not been a problem. Instead of "product_id", orderline id was used to set the relation between the table and the fact table. Numeric field transformations were made as it had been done in customer table.

9.5.3 Time Table

Time dimension was created for the period the data collected, from January 2000 to March 2000. The time dimension field was created using a spreadsheet solution. Spreadsheet solution's date functions were employed to generate date properties such as day of week, week of month.

Time zone differences problem was noticed while creating the time dimension. The time zone differences within and outside USA were dismissed. Alternative solution was to convert the time zone to a constant zone. But this did not seen as a crucial requirement for the analysis.

The day properties were determined from the calendar for the specific time period. Public holidays were marked. Religious days and state specific days were dismissed as they have no standard. The weekend and workday properties of day were kept in the same field.

9.5.4 The Fact Table

In order to provide OLAP analysis a fact table was created to aggregate measures. The numeric fields in the fact table were selected from the orders and clicks tables. The fact table data was designed in orderline level. Each case represented an orderline in table. Measures representing customer shopping patterns were selected according to our business goals. To set up the relation with other tables, foreign keys such as customer_id were placed in the fact table.

The clicks, duration, process time columns were derived from clicks data using transactional SQL commands. The commands can be examined in Appendix C. These fields contained the details of session information before the user had placed the order. The explanations of the derived columns as follows.

Clicks: The number of clicks before the user had placed the order.

Duration: The time in minutes that the user had spent on the site before he/she had placed the order.

Process Time: The time in milliseconds that the user had to wait before his/her clicks were processed by the server.

9.6 OLAP Model and Cube Analysis

OLAP model was constructed in Microsoft Analysis Services. Open Database Connectivity (ODBC) was used to access data from Microsoft Analysis Services. An OLAP cube was designed to analyze our decision problems with the following measures and dimensions.

9.6.1 Measures

Measures were selected including derived variables from the fact table created before. Some of the measures were in order level, which could be interpreted as the sum of orderline levels. All of the variables that could be aggregated as a measure, were selected despite they were not used in the analysis. The list of the measures can be seen in Figure 9-3

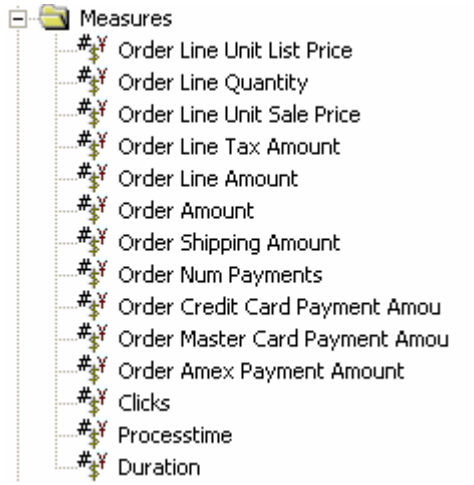


Figure 9-3 Measures Selected for the OLAP cube analysis

9.6.2 Dimensions

Customer Dimension was designed with the concept hierarchies below.

All➔Country➔US State➔CustomerID

For instance as seen in the shared dimension editor of the Microsoft Analysis Services, customer with the id number 23570 belongs to city Anchorage, state AK, country United States.

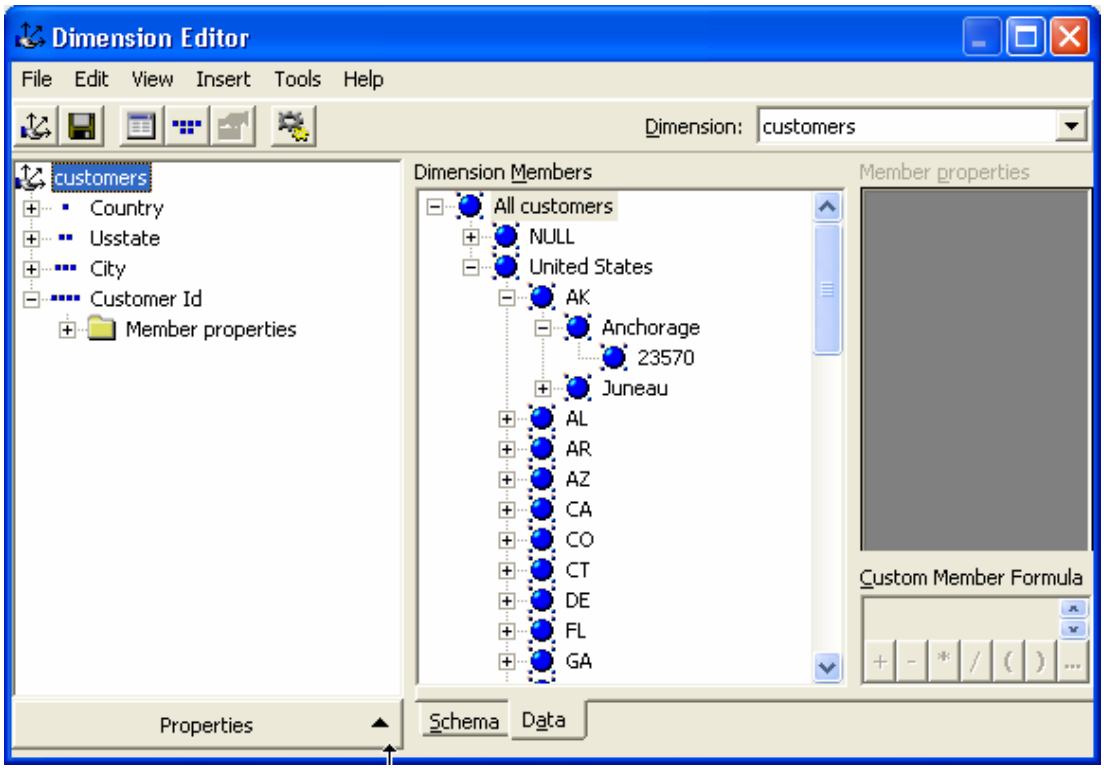


Figure 9-4 The Customer Dimension Customer/ Order Analysis OLAP Cube

Product Dimension was designed with the concept hierarchies below.

Product Level1 Path➔ Product Level2 Path➔ Product Level3 Path

For instance as seen in the shared dimension editor of the Microsoft Analysis Services, BD-HairDiet Spray belongs to BioDepless, Legcare.

Time Dimension was designed with the concept hierarchies below.

Year➔Quarter➔Month➔Day

For instance as seen in the shared dimension editor of the Microsoft Analysis Serverin Figure 9-6 January belongs to quarter1 of year 2000.

9.6.3 OLAP Cube Analysis

The descriptive usage of the OLAP cube can answer business questions like "when?, what? ,where?, which? " After the cube was built it was possible to make descriptive analysis based on product, time and customer dimensions. In our cube it is possible to track the measures like order amount, order quantity, duration on the website before purchase. The measures can be explored according to the dimensions time, product and customer.

Sample questions relevant to our analysis were answered by slice, drill down, roll up operations of the Microsoft Analysis ServerCube Explorer.

- Sample Question 1:

How many leg wear products had we sold to customers in the state CA/United States and what was our revenue in February 2000 and how did it change in March 2000?

Answering such a question makes it easy for the decision maker to facilitate targeted marketing activities. Such a question can also be a starting point to examine the sales trend.

- Sample Question 2:

Can we see our sale amounts according to the estimated income code of our customers?

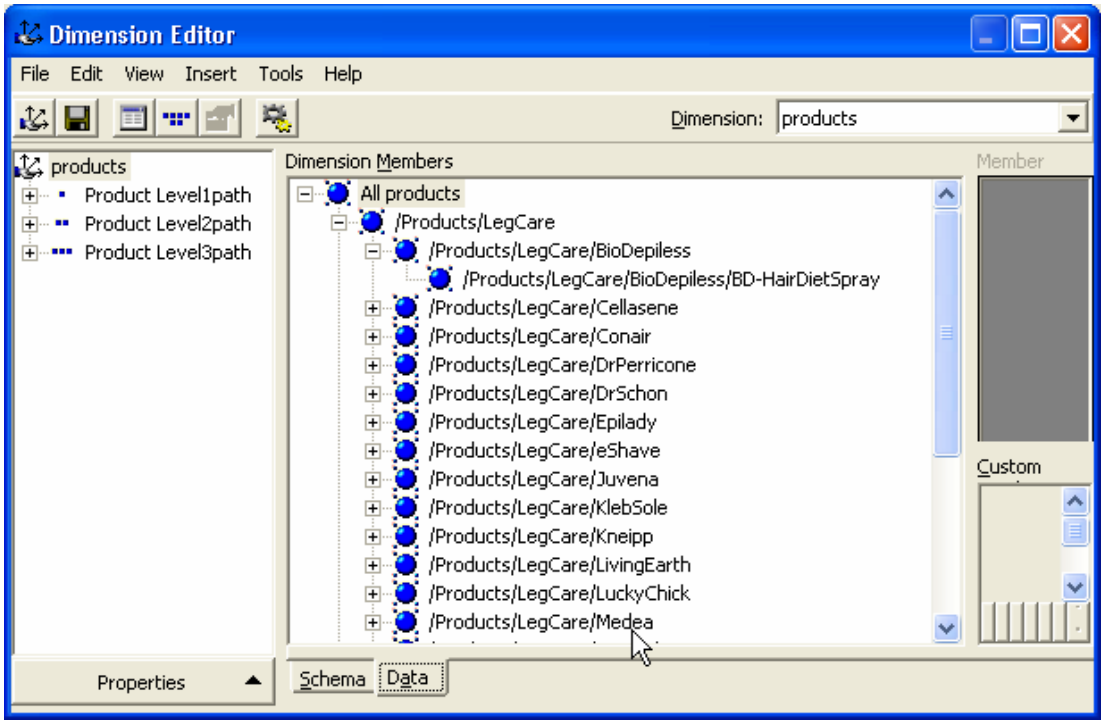


Figure 9-5 The Product Dimension in Customer/ Order Analysis OLAP Cube

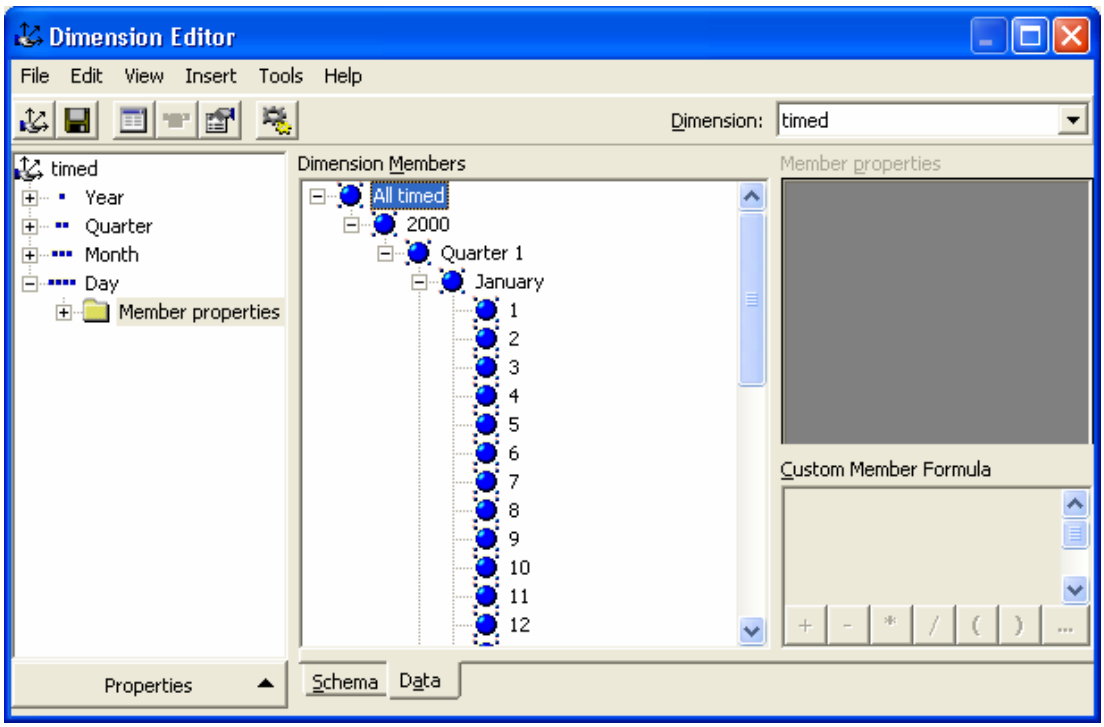


Figure 9-6 The Time Dimension in Customer/ Order Analysis OLAP Cube.

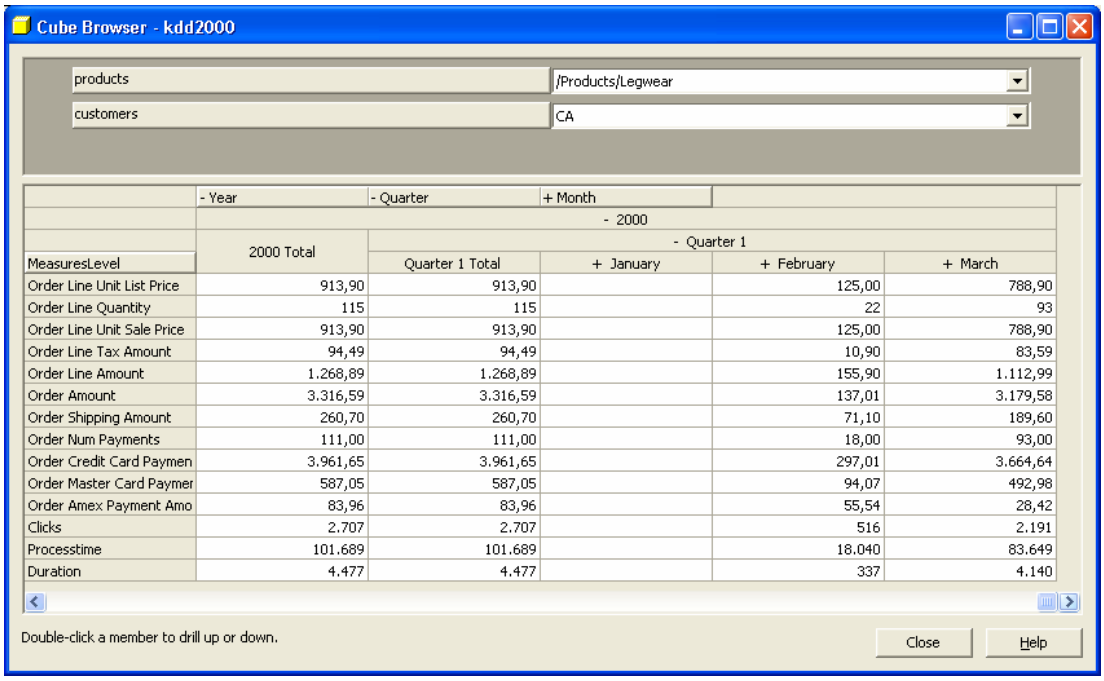


Figure 9-7 OLAP Analysis for Sample Question 1

Answer To Sample Question 1:

- Slice Customer Dimension to United States /CA.
- Slice Products Dimension to Products/Legwear
- Drill down Time dimension to Quarter 1
- As seen the in the output screen of Microsoft Analysis ServerCube Browser in Figure 9-7
- Order Line Quantity in February: 22
- Order Line Quantity in March: 93
- Revenue/Order Amount in February:137,01 $
- Revenue/Order Amount in March:3.179,5 $

Answer To Sample Question 2:

To answer this question we had to create a new virtual dimension based on member properties of customerid in the customer dimension. After the new shared dimension created, it was added to the cube and the cube was reprocessed.

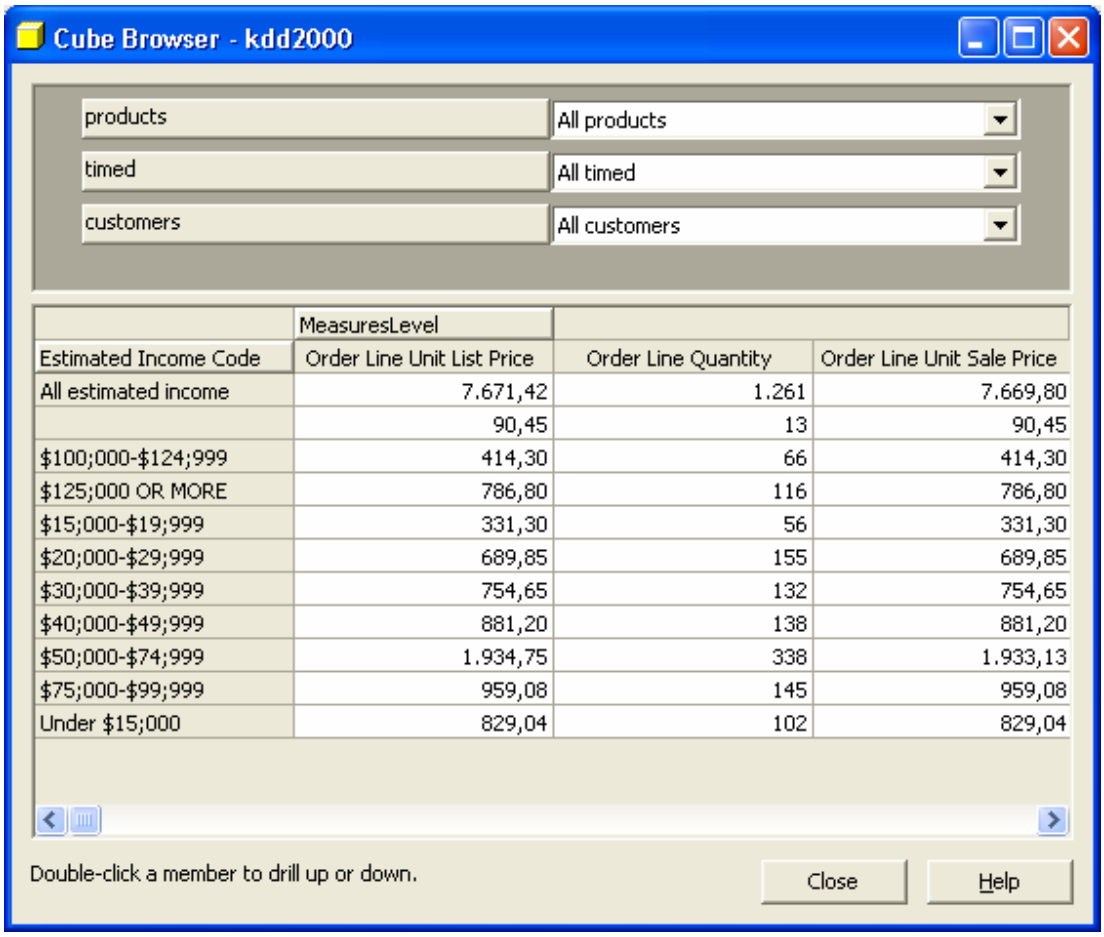


Figure 9-8 OLAP Analysis for Sample Question 2

As seen in the cube browser most of our ordering customers are in the group of estimated income level 50000-75000.

The same analysis could be done by using a third party tool DBMiner which can read OLAP cubes from Microsoft Analysis Server. DBMiner can visualize the cube with three dimensions and two measures. Income Code of customers against measures; order line amount, order line quantity was shown in the Figure 9-9. The lightest colored cube refers the income group which had the biggest order line quantity. The biggest cube in size points out the income group which had the largest order amount. The options of the visualization can be set through the menu (figure) in order to extend the analysis.

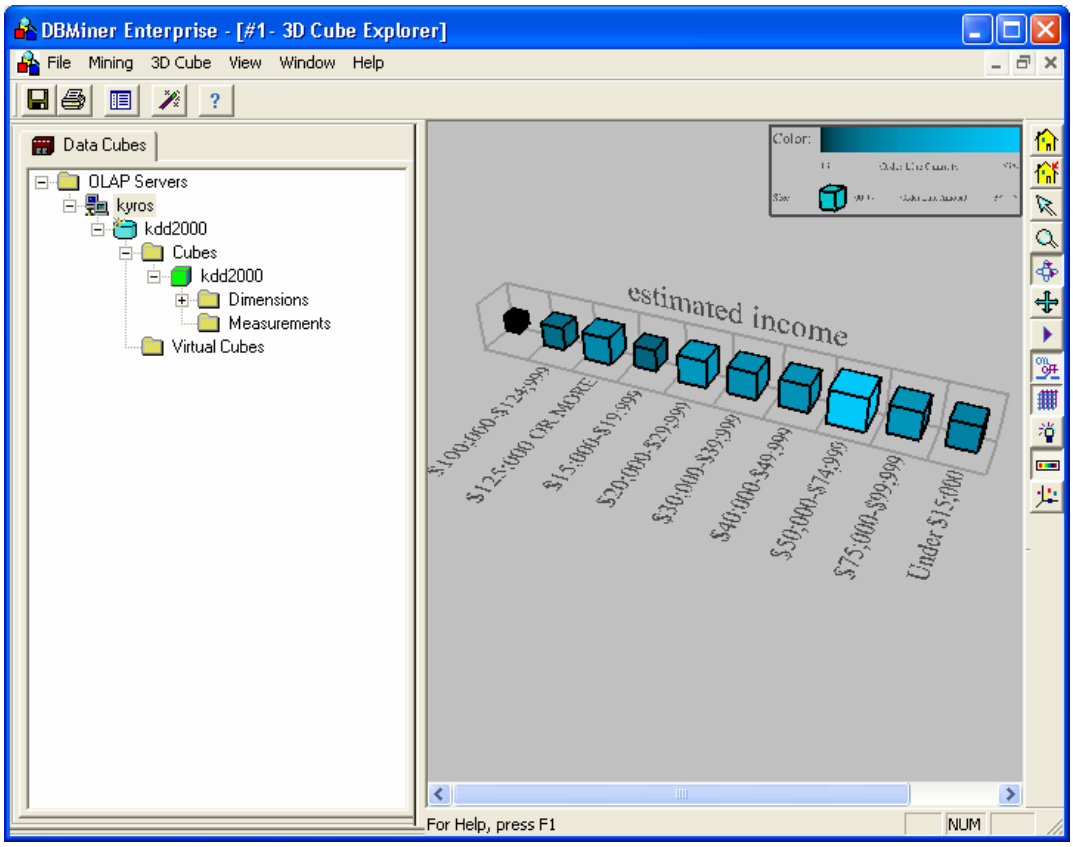


Figure 9-9 3D Cube Visualization from DBminer Software.

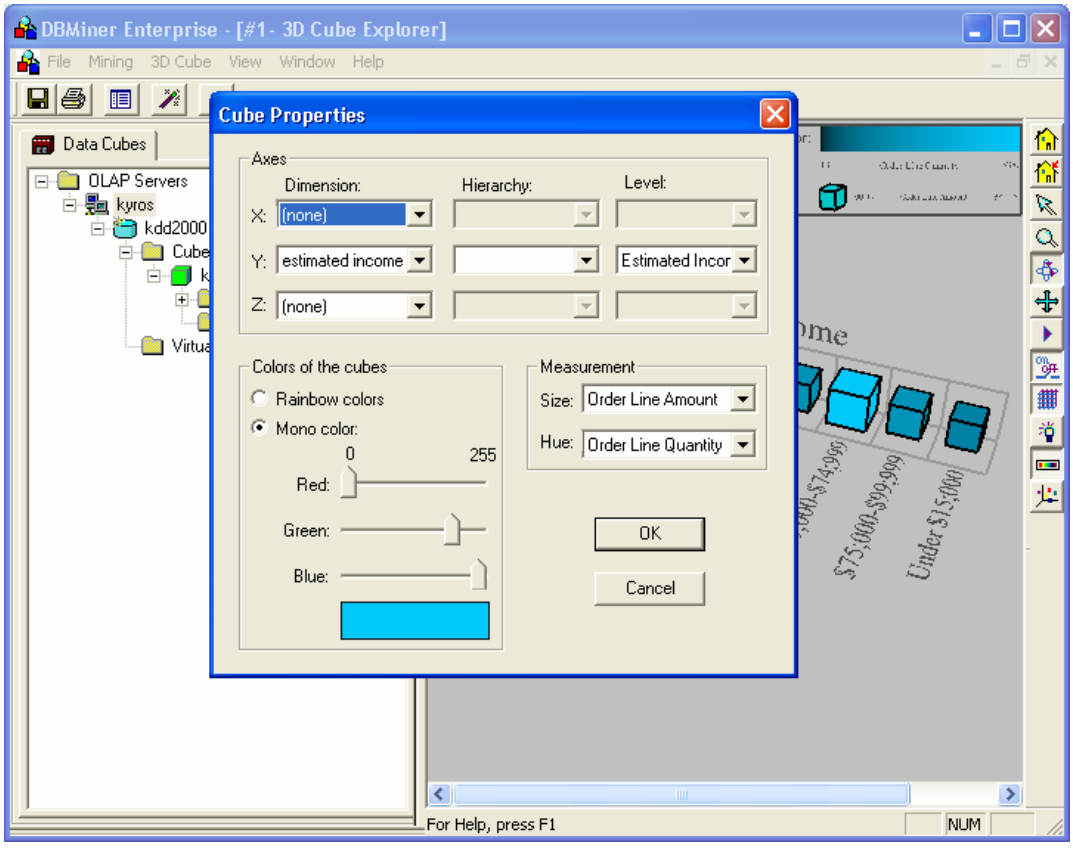


Figure 9-10 The Settings of 3D Cube Visualization in DBMiner.

9.7 Determining Valuable Customers

The objective of the models is to classify and predict the customers who spent more than 12$ average per order. The business objective was to identify or predict valuable customers. Such a characterization or classification can facilitate targeted marketing activities as mentioned in detail in previous chapters.The promotions and direct marketing activities can be customized to the characteristics of the customers.

Two different models were built with different input variables. Microsoft Decision Tree Mining Model and SPSS Answer Tree were used. The results were evaluated separately.

9.7.1 Microsoft Decision Tree Algorithm

Microsoft decision tree algorithm was used to build the first model. The algorithm can handle both quantitative and qualitative data. There were two data source options for Microsoft Decision Tree, OLAP or Relational model. OLAP data source required a cube and its advantage was to calculate aggregations from the cube. Selecting OLAP data source would also let us skip the step of setting table relationships manually. Relational Model's advantage was that it did not require a cube, relationships could be set manually and more robust to errors as we tested both in our model.

We used the relational model in our prediction model over modified version of orders table. The input variables for the model are shown in Figure 9-11. The variables marked with diamonds were the predicted variables. A variable could be both chosen input and predictable in the tool we had used.

Figure 9-11 Input variables for Microsoft Decision Trees Model

The resulting tree was as follows. The hue of the boxes refers to the intensity of cases.

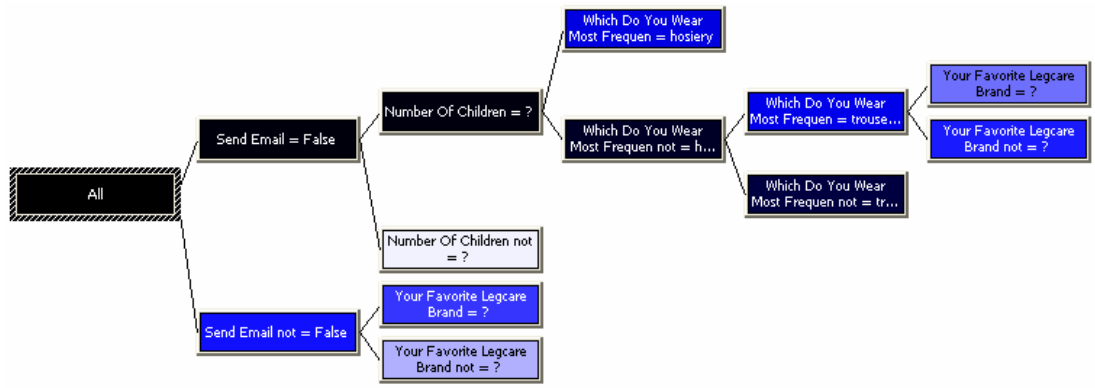


Figure 9-12 Resulting Decision Tree Model

The same tree could be visualized for the customers who spent more than 12 $ average per order as in Figure 9-12. The dark boxes represent the intensity of the targeted variable.

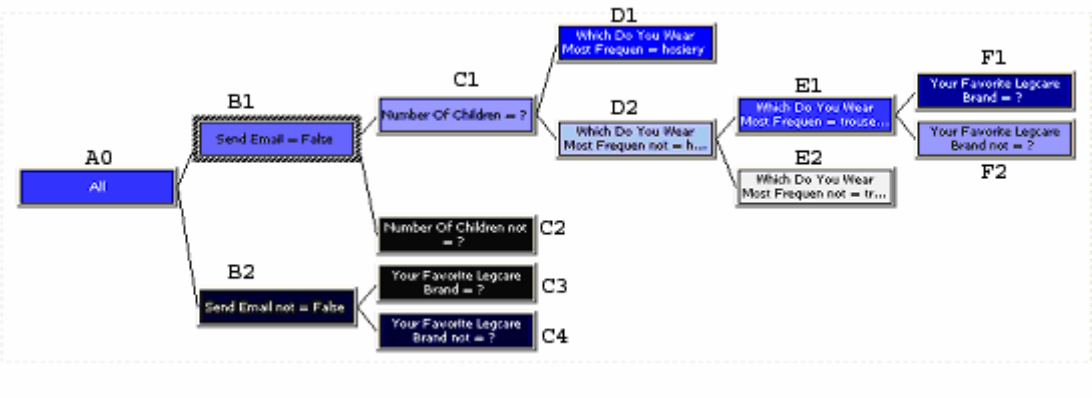


Figure 9-13 Resulting Decision Tree Visualized by the Target Variable

Details of the analysis were summarized in table.

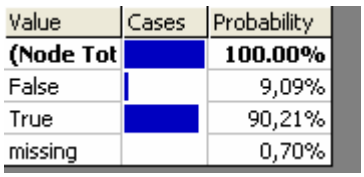
Table 9-1 Microsoft Decision Tree Model Summary

Following Node / Separator	Value	Cases	Probability	Histogram
A0 ALL	**(Tree Tol**	**1831**	**100.00%**	**100.00%**
	False	1407	76,84%	76,84%
	True	424	23,16%	23,16%
	missing	0	0,00%	0,00%
B1-Send Email=False A0	**(Node To**	**1641**	**100.00%**	**100.00%**
	False	1380	84,00%	84,00%
	True	261	15,94%	15,94%
	missing	0	0,06%	0,06%
B2- Send Email not=False A0	**(Node To**	**190**	**100.00%**	**100.00%**
	False	27	14,51%	14,51%
	True	163	84,97%	84,97%
	missing	0	0,52%	0,52%
C1-Number of Children=? B1	**(Node To**	**1612**	**100.00%**	**100.00%**
	False	1379	85,45%	85,45%
	True	233	14,49%	14,49%
	missing	0	0,06%	0,06%
C2-Number of Children not=? B1	**(Node To**	**29**	**100.00%**	**100.00%**
	False	1	6,25%	6,25%
	True	28	90,62%	90,62%
	missing	0	3,12%	3,12%

C3-Your favorite Leg Care Brand=?

B2

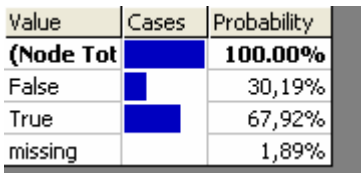
Value	Cases	Probability
(Node Tot	**140**	**100.00%**
False	12	9,09%
True	128	90,21%
missing	0	0,70%



C4- Your favorite Leg Care Brand Not=?

B2

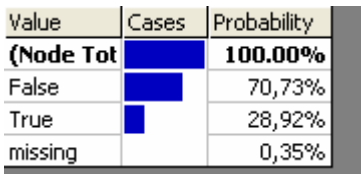
Value	Cases	Probability
(Node Tot	**50**	**100.00%**
False	15	30,19%
True	35	67,92%
missing	0	1,89%



D1-Which do you wear most frequent=hosiery

C1

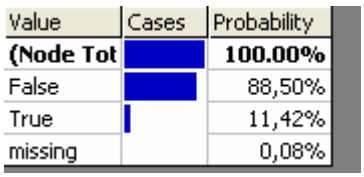
Value	Cases	Probability
(Node Tot	**284**	**100.00%**
False	202	70,73%
True	82	28,92%
missing	0	0,35%



D2- Which do you wear most frequent not=hosiery

C1

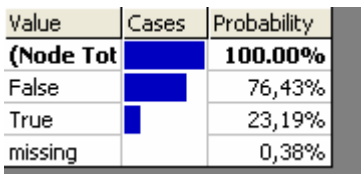
Value	Cases	Probability
(Node Tot	**1328**	**100.00%**
False	1177	88,50%
True	151	11,42%
missing	0	0,08%



E1- Which do you wear most frequent =trouser socks

D2

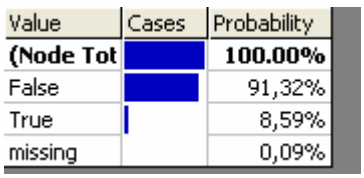
Value	Cases	Probability
(Node Tot	**260**	**100.00%**
False	200	76,43%
True	60	23,19%
missing	0	0,38%



E2- Which do you wear most frequent not=trouser socks

D2

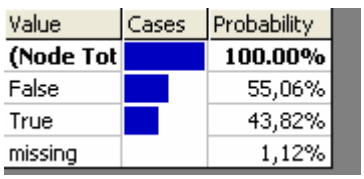
Value	Cases	Probability
(Node Tot	**1068**	**100.00%**
False	977	91,32%
True	91	8,59%
missing	0	0,09%



F1-Your Favorite Legcare Brand=?

E1

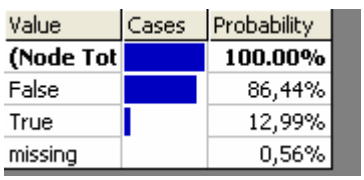
Value	Cases	Probability
(Node Tot	**86**	**100.00%**
False	48	55,06%
True	38	43,82%
missing	0	1,12%



F2- Your Favorite Legcare Brand not=?

E1

Value	Cases	Probability
(Node Tot	**174**	**100.00%**
False	152	86,44%
True	22	12,99%
missing	0	0,56%



9.7.2 Evaluation of the Results

The first branch (B1/B2) was separated by the variable “send email”. 190 cases over 1831 customers were contacted via email. 84.97% of these cases spend over 12$ average per order. (B2) Those were not contacted via email had a ratio of 15.94% for the same target variable (B1). The strong distinction in this branch showed us sending email to customers made a positive impact to sales over 12$ per order.

The distinction had been made for the customers whom had not been contacted via email in node B2 by two branches. The tree was divided into two branches C1 and C2 in this point by the variable “Number of children”. “Number of Children=?” refers to null values in the database for this field. 85.45% of customers responded to false to our target variable in this node (C1).The other branch “Number of Children not=?” refers to values other than null(C2). The probability of “true” responders to our target variable was 90.62%. The tree stopped in this node(Terminal node). This branch could inform the decision maker that the customers whom had children had a potential of profitable purchases in the retail shop.

The distinction had been made for the customer whom had been contacted via email with the in node B1 by two branches. The tree was divided into two branches in this point C3 and C4 by the variable “Your favorite Leg Care Brand”. Responders “Your favorite Leg Care Brand=?” to null values. 90.21% percent of the customers in this branch had purchased over 12$ per order(C3). The other branch at this point had the same probability of 67.92 %.(C4) The probabilities were not as they were expected to be. By heuristics, the customers who responded the question were expected to be in the group of purchasers over 12$ average per order. Although the probability was not so low, the decision maker should avoid concluding strong results leaning on this branch.

C1 was divided into two branches with the variable “Which do you wear most frequent”. The customers who responded “hosiery” had the probability of 70.73% false for the target variable (D1). Who responded other than hosiery had the same probability with the value 88.50% (D2).

D2 is divided into two branches with the variable “Which do you wear most frequent”. The customers who answered “trouser socks” to this question 76.43%(False) for the target variable. The customers who answered other than

trouser socks had the probability of 91.32 % (False) for the target variable. This strong probability could easily tell us that customer who wears “trouser socks” has the potential of being profitable customer. The decision maker can use this information while setting the promotion and direct marketing activities.

In the last branch of the tree we see the variable “Your Favorite Legcare Brand” dividing E1 into two branches. Customers whom did not respond to this question had the probability of 55.06% (False) for the target variable. Responders to this question had the probability of 86.44% (False). This may be clue for the analyzer that some of the popular brands were not in the assortments of the retailer. Decision maker should need more information to make strong conclusions on these probabilities.

9.7.3 SPSS Answer Tree Classification Model

SPSS Answer Tree package was used to build the second model. Exhaustive CHAID algorithm was used to grow the decision tree. The algorithm uses F and Chi-Square statistics to select predictors. Each split can have several nodes.

Data was imported via ODBC using SPSS. Necessary data transformations were made for data preparation. Ordinal variables were recoded using SPSS data transformation functions.

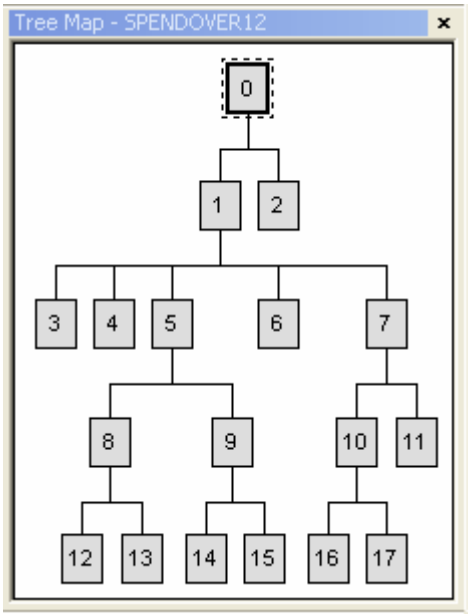
The input variables and their measurement scales were;

- YOURFAVO: Your Favorite Legcare Brand /Nominal
- WHICHDOY: Which Do You Wear Most Frequent/Nominal
- HOWDOYOU: How Do You Dress For Work/Nominal
- YEARHOME: Year Home was Bought/Scale
- NUMBER1:Number Of Vehicles/Scale
- GENDER: Gender Of Customer/Nominal
- OTHERI1:Gender Of Other Individual Purchase Made For/ Nominal
- MARITALS: Marital Status/Nominal
- HOWMNYP: How May Pairs Do You Purchase/Ordinal
- SENEMA: Send Email/Nominal

- HOWOFDYP: How OftenDo You Purchase/ Ordinal
- HOMEMARV: Home Market Value/Ordinal
- ESTINC4: Estimated Income Code/Ordinal

The map of the resulting tree was shown in Table 9-2.

Table 9-2 SPSS Answer Tree Map View



Nodes Summary

Total number of nodes:18

Total number of levels:4

Total number of terminal nodes:11

Stopping Rules

Max Tree Depth: Limits the levels of the tree.

It was set to 7.

Minimum Number Of Cases For Parent Node:40

Minimum Number Of Cases For Child Node:20

Splitter variables and nodes are summarized in Table 9-3.

Table 9-3 Variables and Nodes

Level	Splits	Variable	Split Node
1	2	Send Email/Nominal	0
2	5	Which Do You Wear Most Frequent /Nominal	1
3	2	How Often Do You Purchase	5
3	2	Your Favorite Legcare Brand	7
4	2	Your Favorite Legcare Brand	8
4	2	Your Favorite Legcare Brand	9
4	2	Marital Status	10

The summary of probabilities in nodes and number of branches in levels are summarized in Table 9-4.

Table 9-4 SPSS Answer Tree Solution Summary

Level 0

Node 0		
Category	%	n
True	17,95	243
False	82,05	1111
Total	(100,00)	1354

Level 1

Node 1		
Category	%	n
True	13,83	176
False	86,17	1097
Total	(94,02)	1273

Node 2		
Category	%	n
True	82,72	67
False	17,28	14
Total	(5,98)	81

Level 2

Node 3		
Category	%	n
True	3,42	8
False	96,58	226
Total	(17,28)	234

Node 4		
Category	%	n
True	13,76	30
False	86,24	188
Total	(16,10)	218

Node 5		
Category	%	n
True	28,02	65
False	71,98	167
Total	(17,13)	232

Node 6		
Category	%	n
True	6,61	25
False	93,39	353
Total	(27,92)	378

Node 7		
Category	%	n
True	22,75	48
False	77,25	163
Total	(15,58)	211

Level 3

Node 8		
Category	%	n
True	47,73	42
False	52,27	46
Total	(6,50)	88

Node 9		
Category	%	n
True	15,97	23
False	84,03	121
Total	(10,64)	144

Node 10		
Category	%	n
True	8,40	10
False	91,60	109
Total	(8,79)	119

Node 11		
Category	%	n
True	41,30	38
False	58,70	54
Total	(6,79)	92

Level 4

Node 12		
Category	%	n
True	10,34	3
False	89,66	26
Total	(2,14)	29

Node 13		
Category	%	n
True	66,10	39
False	33,90	20
Total	(4,36)	59

Node 14		
Category	%	n
True	25,56	23
False	74,44	67
Total	(6,65)	90

Node 15		
Category	%	n
True	0,00	0
False	100,00	54
Total	(3,99)	54

Node 16		
Category	%	n
True	2,53	2
False	97,47	77
Total	(5,83)	79

Node 17		
Category	%	n
True	20,00	8
False	80,00	32
Total	(2,95)	40

Misclassification Matrix				
		Actual Category		
		True	False	Total
Predicted Category	True	106	34	140
	False	137	1077	1214
	Total	243	1111	1354

	Risk Statistics
Risk Estimate	0,126292
SE of Risk Estimate	0,00902739

Figure 9-14 SPSS Answer Tree Misclassification Matrix

9.7.4 Evaluation of the Results

First level, Node 0, had the probabilities 17.95(F), 82.05(T) to the target variable "spent over 12$ average per order". This node was divided into two branches Node1 and Node 2 by the variable "send email". The customers whom were contacted via email had the probability of 82.72. This finding supports the results of the Microsoft Classification Tree. The decision maker should consider the strong probability that sending email strengthens the probability of purchases over 12$ per order.

Node 1 was divided into five branches by the variable "which do you wear most frequent". Nodes 3, Node 4, Node 6 were terminal nodes. The terminal nodes, Node 3 and Node 4 had the probability of 96.58(F) and 86.24(F) for the target variable respectively. Node 3 was represented by the variable "athletic socks" while Node 4 was represented by null values. Node 6 had the probability of 93.39(F) for the target variable which was represented by the variable "casual socks". The decision maker may conclude that the certain leg wear products are not bought by customers spent more then 124 per average.

Node 7 was divided into two nodes Node 10 and Node 11 by the variable "your favorite legcare brand" which resulted with the terminal node, Node 11. Node 11 had the probability of 41.30(T) for the target variable. The brands representing this node can be considered as the motivators to spend more than 12$ per order average.

An interesting path in the tree was Node 8, which was separated by the variable "how often do you purchase". As expected the customers who purchased often had higher probability of spending more than 12$ per average 47.73 (T). This node was split by the variable "your favorite legcare brand" and resulted with a terminal node, node, Node 13/66.10 (T).

The misclassification matrix in Figure 9-14 shows that the model's estimation risk was 0.12 approximately. The model estimated only 42% of customers who spent over 12$, correctly. The model's accuracy of estimating the customers who spent more than 12$ per average was low. The model was more accurate on estimating the customers who did not spend more than 12$ on average (probability was 89%). The matrix summarizes the estimation accuracy of the tree. We can find out the reliability of the tree by interpreting the results.

According to the misclassification matrix, our model was not so accurate and the results were not reliable in overall. On the other hand some terminal nodes were pointed considerable finding about the characteristics of customers who spent more than 12$ per order on average.

9.7.5 Problems and Considerations with the Analysis

The huge number of missing values weakened this analysis's results. The dataset, orders data, was not sufficient to make strong interpretations on results despite it gave a good starting point.

Some of the string nominal variables could have been transformed to numeric ordinal ones in order to support the algorithm. Such an effort had not been made because the processing principles of the algorithm were not known.

In the second model string and nominal variables were transformed to numeric and ordinal variables in SPSS.

The missing values in the dataset except the numerical ones were left as they were. They could have been replaced by "null" values in SQL Server. However this problem did not affected to the analysis but interpretation of the results.

9.8 Customer Segmentation

The objective of the model is to find out the key variables that cluster the customers. The segments after clustering can be used in various targeted marketing activities such as personalized promotions.

Microsoft Clustering Algorithm in Microsoft Analysis Server was used as a tool to construct the model. The algorithm groups the cases according to the similarity of the variables, but the technique is not known. As it is in the decision tree algorithm of the same tool there is no limitation on variable types. The user can choose either OLAP or relational model as data source. The edit screen of the model is as in the Figure 9-15

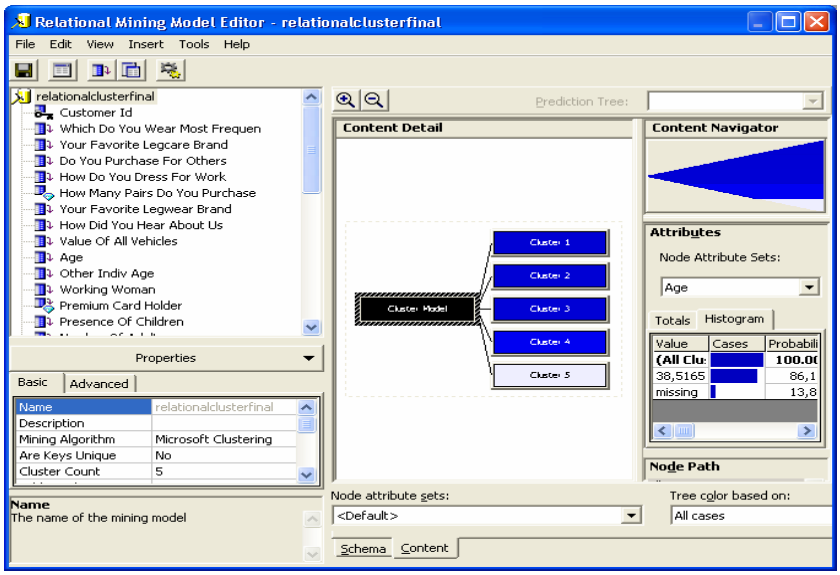


Figure 9-15 Microsoft Clustering Model in Analysis Services

Cluster count was set to 5 for the model. The details of the analysis and the properties of the nodes were summarized in the Table 9-5. The hue of the nodes darken as the intensity of the selected variable increases.

We may conclude the following facts about the clusters by exploring the tree view of analysis services.

98.26% of customers in cluster 2 were mail order buyers while 72.74 % were mail order buyer at the top level.

98.50% of customers in cluster 4 had no finance retail activity 89.90 had no finance retail activity at the top level.

92.51% of customers in Cluster 2 and 81.23% of customers were married while 63.21 were married at the top level.

60.27% of customers in Cluster 2 were working woman while 36.12% of the cases were same so at the top level.

Berkshire and Donna Kran were favorite brands among customers in Cluster 5 with the probabilities 17.19% and 16.70%. The same brands have the ratios 2.01% and 5.18% at the top level respectively.

Similar findings could be generated to serve decision maker. The Clusters can be used to predict the customer preferences and retail patterns.

Table 9-5 Microsoft Clustering Algorithm Summary of Results

CLUSTER1

Occupation = SELF EMPLOYED SALES/MARKETING ,
Clicks = 65 ,
Clicks = 104 ,
Clicks = 88 ,
Last Retail Date = 1994-10-31 ,
Clicks = 97 ,
Clicks = 136 ,
Last Retail Date = 1994-07-31 ,
Clicks = 76 ,
Last Retail Date = 1995-11-30 ,
Other Indiv Age = 72 ,
Last Retail Date = 1992-01-31 ,
Clicks = 53 ,
Other Indiv Age = 84 ,
Last Retail Date = 1996-04-30 ,
Clicks = 102 ,
Number Of Adultscon = 0 ,
Other Indiv Age = 94 ,
How Did You Hear About Us = ,
Gender = NULL

CLUSTER2

Home Market Value = $775;000-$999;999 ,
Last Retail Date = 1993-11-30 ,
Clicks = 52 ,
Last Retail Date = 1994-02-28 ,
Clicks = 143 ,
Other Indiv Age = 54 ,
Other Indiv Gender = Male ,
Other Indiv Occupation = PROFESSIONAL/TECHNICAL ,
Other Indiv Age = 46 ,
Household Status = NAME APPEARING ON INPUT IS INDIVIDUAL 2 ,
Working Woman = True ,
Marital Status = Married ,
Last Retail Date = 1996-01-31 ,
Premium Card Holder = True ,
Mail Order Buyer = True ,
Other Indiv Occupation = SALES/SERVICE ,
Other Indiv Occupation = ADMINISTRATIVE/MANAGERIAL ,
Gender = Female ,
Last Retail Date = 1997-06-30 ,
Home Market Value = $50;000-$74;999

CLUSTER3

Last Retail Date = 1993-08-31 ,
Other Indiv Age = 20 ,
Clicks = 40 ,
Last Retail Date = 1994-11-30 ,
Last Retail Date = 1992-09-30 ,
Clicks = 8 ,
Clicks = 67 ,
Last Retail Date = 1994-08-31 ,
Last Retail Date = 1992-08-31 ,
Clicks = 59 ,
Other Indiv Occupation = SELF EMPLOYED ,
Clicks = 51 ,
Other Indiv Gender = Male ,
Household Status = NAME APPEARING ON INPUT IS INDIVIDUAL 2 ,
Other Indiv Age = 36 ,
Number Of Adultscon = 2 ,
22,8450761998008 <= Age <= 45,4155810235244 ,
Other Indiv Age = 32 ,
Other Indiv Age = 34 ,
Other Indiv Occupation = SELF EMPLOYED MANAGEMENT

CLUSTER4

Other Indiv Occupation = SELF EMPLOYED RETIRED ,
Last Retail Date = 1995-12-31 ,
Clicks = 78 ,
Duration <= 1 ,
Your Favorite Legwear Brand = ,
Age > 81,444109194218 ,
Home Market Value = $350;000-$399;999 ,
Your Favorite Legwear Brand = Berkshire ,
Other Indiv Age = 58 ,
Spendover12perorder = True ,
Last Retail Date = 1996-07-31 ,
Number Of Adultscon = 1 ,
Estimated Income Code = Under $15;000 ,
Number Of Vehicles = 1 ,
Marital Status = Single ,
Household Status = NAME APPEARING ON INPUT IS INDIVIDUAL 1 ,
Last Retail Date = 1997-04-30 ,
Home Market Value = $225;000-$249;999 ,
Presence Of Children = False ,
Other Indiv Age =

CLUSTER5

Clicks = 54 ,
Clicks = 61 ,
Last Retail Date = 1995-01-31 ,
Number Of Adultscon = 1 ,
Marital Status = Inferred Single ,
Household Status = NAME APPEARING ON INPUT IS INDIVIDUAL 1 ,
Other Indiv Age = 22 ,
Own Or Rent Home = NULL ,
Mail Order Buyer = False ,
Other Indiv Age = ,
Other Indiv Gender = NULL ,
Your Favorite Legwear Brand = Greg Norman ,
Presence Of Children = False ,
Home Market Value = ,
Own Or Rent Home = Renter ,
Last Retail Date = ,
Other Indiv Occupation = NULL ,
Marital Status = NULL ,
0 <= Length Of Residence <= 8,2591409229223 ,
Estimated Income Code = $20;000-$29;999

Cluster1

Value	Cases	Probability
(Cluste	**150,68**	**100.00%**
25775,6	63,10	41,88%
missing	87,58	58,12%

Cluster2

Value	Cases	Probability
(Cluste	**150,12**	**100.00%**
18616,9	84,15	56,06%
missing	65,97	43,94%

Cluster3

Value	Cases	Probability
(Cluste	**149,20**	**100.00%**
18280,6	78,65	52,71%
missing	70,55	47,29%

Cluster4

Value	Cases	Probability
(Cluste	**124,07**	**100.00%**
8613,52	33,00	26,60%
missing	91,07	73,40%

Cluster5

Value	Cases	Probability
(Cluste	**23,92**	**100.00%**
14824,2	9,10	38,04%
missing	14,82	61,96%

CONCLUSION

The internet has provided many opportunities for commercial activities. E-commerce on the internet is increasing its volume day by day. Understanding and profiling customer behavior on the internet is crucial for internet retailers. Web is a flexible environment and it can almost be customized for every single user. Web is an extremely convenient environment to capture user interaction data and transform it directly to the warehouse by automated tools. The exciting point is here to use this data to modify products, services and the shape of the store according to customer's needs. Extraction of knowledge, behind the interaction between user and the web site can be employed to data mining process.

First thing we found out was the importance of data quality. It was very hard to make a valid analysis on data with many missing values and less trustable cases. Data cleaning and preparation took most of our time and effort. We concluded that the most important step of web mining is to set up a well designed data accumulation system based on business goals. The purpose of such a system should be to collect the ready to be analyzed data. The data collected without this perspective increases storage costs, limits the tools that can be used with data and hardens to extract cost effective knowledge.

Another observation about the tools we had used was that data mining software's user interface and functionality was not satisfactory. It is hard to develop a generic data mining tool with a user friendly interface. More specific tools can be developed for specific data structures and purposes. SPSS Answer Tree is a good example of a dedicated analysis tool but it yet requires expert usage and special data format.

Data mining is a hopeful technology for many areas dealing with mass data. But the mining infrastructure, selection and collection data plays a crucial role to get worthwhile results. There are varieties of tools using varieties of algorithms. Every tool and algorithm has strengths and weaknesses over different data types and

analysis. It is extremely possible to get turn down with meaningless results if wrong combinations were selected.

The process of the data mining is not automated; it requires patience and careful analysis of the user to notice extraordinary patterns. It requires making iterations to the beginning at the final stage even the infrastructure was designed well.

It is obvious that the business era is evolving rapidly to more customer centric orbit. Products are getting more customized; manufacturing systems are getting more flexible. New strategic management approaches like customer relationship management are appearing in order to sustain competitive advantage. Products are not valued by direct manufacturing costs; they are valued by the knowledge behind them.

Data mining technology enables core benefits for enterprises. Data mining facilitate statistical and analytical analysis of data, yet it achieves beyond statistical analysis. Data mining technology has the ability of extracting hidden patterns beyond data which can not be become apparent by heuristic approaches. The descriptive analysis enabled by data mining technology and OLAP solve the mass data exploration problem.

It is seen from outside the window that, data mining packages in the market are magic wands which explodes business. Enterprises should act carefully when making investments on data mining software packages. This is because data mining is an emerging technology, it is hard to apply, it has many requirements before it can be applied efficiently. The enterprise must have set a strong e-business structure and a healthy data accumulation structure based upon it. The driller is important when mining but if you have nothing to extract, it won't worth your effort. Defining business problems and decision support needs should be the first step in data mining process. The role of selecting and accumulating knowledge rich data is crucial. Another problem which the data mining software users face is the complexity of data mining packages. Not only they require dedicated warehouses, they are hard to operate. They do not use standard algorithms and the most important, they are not user friendly. Developing industry specific data warehouse designs and mining algorithms would increase the benefits of data mining for enterprises.

BIBLIOGRAPHY

Berry, M. J. A.; Linoff, G,. 1997, Data Mining Techniques. For Marketing, Sales and Customer, Support. Wiley Computer Publishing

Berry, Michael J. and A. Linoff, Gordon., 2000, Mastering Data Mining : The Art and Science of Customer Relationship Management ,Wiley Computer Pub., New York

Blaxton, T. and Westphall, C., 1998, Data Mining Solutions; Methods and Tools for Solving Real-World Problems, Willey Computer Publishing, New York.

Brainerd, Jeffrey, Becker, Berry, 2001, Case Study: E-Commerce Clickstream Visualization*, Proceedings of the IEEE Symposium on Information Visualization 2001* , (INFOVIS'01)

Chen, Lin, Zheng, Lin, Fan, Liu, Yin, Ying, Eei, Wenyin, Liu, 2002,User Intention Modeling in Web Applications Using Data Mining, *World Wide Web: Internet and Web Information Systems*, **5**, 181–191

Chiu, Chao-Min, 2001, Towards Integrating Hypermedia and Information Systems on the Web, *Information & Management* , 1980 (2002) 1–11

Dahlan, N., Ramayah, T., Hoe A.K., 2002**,** Data Mining in the Banking Industry: an Exploratory Study, *International Conference on Internet Economy And Business*

Doherty, Patricia, 2000 ,Web Mining The E-Tailer's Holy Grail, www.dmreview.com/editorial/dmreview/print_action.cfm?EdID=1891

Feeders, A., Daniels, H., Holsheimer M., 2000 , Methodological and Practical Aspects of Data Mining, *Information & Management,* **37**, 271-281

Greening, Dan R., 2000, "Data Mining on the Web", www.webtechniques.com/archives/2000/01/greening/

Grossman R., **Kasif, S., Moore, R., Rocke, D., Ulman, J.,**1998 , Data Mining Research: Opportunities and Challenges , *NSF Workshops on Mining Large Massive and Distributed Data*, January 21.

Groth, Robert, 1999, Data Mining: Building Competitive Advantage, Prentice Hall, New Jersey.

Han, J. and Kamber, M., 2001, Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco

Hand, D., Mannila, H., Smyth, P., 2001, Principles of Data Mining, MIT Press, London.

Hongjun, Lu and Ling, Feng, 1998, Integrating database and World Wide Web Technologies,*World Wide Web*, **1** , 73–86

Hui, S.C, Jha, G., 2000, Application: Data Mining for Customer Service and Support, *Information & Management,* **38,** 1-13

Jennings, Michael F. ,2000, "Using clickstream as an e-Source for the e-Business Intelligence Environment", www.ittoolbox.com/peer/CEBIE.htm

Kimball , R., Reeves, Laura, Ross, Margy, Thornthwaite, Warren,1999, The Data Warehouse Lifecycle Toolkit, Willey Computer Publishing, New York.

Kimball, R. and Merz, Richard, 2000, The Data Webhouse Toolkit, Willey Computer Publishing, New York.

Masand, Brij, Spiliopoulou, Myra, 2000, Web Usage Analysis and User Profiling, *International WEBKDD'99 Workshop San Diego, CA, USA, August 15*,1999 Revised Papers, Springer, New York

Mattison, Rob, 1999,Web Warehousing and Knowledge Management, McGraw-Hill, New York.

McDunn, 2002, "Web Server Log File Analysis – Basics", www-group.slac/stanford.edu/techpubs/logfiles/info.html

Mena, Jesus ,2002, "Integrating and Mining Web Data in Your Warehouse", http://www.dmreview.com/editorial/dmreview/print_action.cfm?EdID =1402

Rud, Olivia Parr, 2001, Data Mining Cookbook; Modeling Data for Marketing, Risk and Customer Relationship Management, Willey Computer Publishing, New York.

Thomas,Khabaza, Thomas, Reinartz, Colin, Shearer, and Wirth Rüdiger, 1996 CRISP-DM 1.0:Cross Cross Industry Standard Process for Data Mining, http://www.crisp-dm.org

Tittel, Ed , 2001,Understanding Web Server Log Files, http:// www.searchsystemsmanagement.techtarget.com/tip/1,289483, sid20_gci849167,00.html

Tiwana, Amrit, 2001, The Essential Guide to Knowledge Management: E-business and CRM Applications, Prentice Hall PTR, Saddle River

Watson, Hugh J., Goodhue, D.L., Wixom, B.H., 2002, The Benefits of Data Warehousing: Why Some Organizations Realize Exceptional Payoffs, *Information & Management*, **39,** 491-502

Wirth , Roger, Hipp, Jochen, 2001, CRISP-DM: Towards a Standard Process Model for Data Mining, http://www-db.informatik.uni-tuebingen.de/ forschung/papers/

Zaiane, Osmar Rachid, 1999, Resource and Knowledge Discovery from the Internet and Multimedia Repositories, *PhD Thesis*, Simon Fraser University, Canada

APPENDIX A :SUMMARY OF PHASES IN CRISP-DM 1.0

Table A.1 Sumary of Phases in CRISP-DM 1.0

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **DetermineData Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion / Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modeling Technique** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

APPENDIX B :KDD CUP 2000 INTRODUCTION AND QUESTIONS

This document contains information confidential to Blue Martini Software and Gazelle.com Inc. Use of the data is restricted by a non-disclosure agreement that you must have signed at http://www.ecn.purdue.edu/KDDCUP/

KDD Cup 2000 Introduction and Questions

Updated 7/14/2000

Change log at end of file

Warning: The organizing committee reserves the right to update this document and modify the evaluation criteria or add restrictions.

Changes to the doc will be highlighted off the web page
http://www.ecn.purdue.edu/KDDCUP/data.
While we would like to give you a stable task and do expect any major changes, it is possible that additional information about the data or specific attributes may render some tasks unfair or unrealistic. This is the main reason why the test set is not being released at this stage. It is possible that we may need to eliminate an attribute or two if it "leaks" the result (e.g., if tax were given, it would predict the purchase amount).

We will acknowledge participants who discover such leaks should they wish to be mentioned.

Introduction

This document describes the questions posed for the KDD Cup competition, the evaluation criteria, and descriptions of the domain, data collection, and file formats. Look at the web page **http://www.ecn.purdue.edu/KDDCUP/data** for additional questions/FAQ.

The competition this year involves web clickstreams and purchase transactions collected by the Blue Martini Software application running at Gazelle.com, which sells legware and legcare products. The data provided was collected directly at the store, with minimal cleaning that is described below. No attempts were made to create a theoretically sound dataset; in fact, we know the data was affected by external factors (e.g., marketing programs) and relevant information is described below.

Questions

Participants will be judged on the following five questions. Evaluations and awards will be given for each question independently, i.e., it is not necessary to submit entries to all questions. Multiple winners will be announced for each question. The first two tasks are classification questions that will be judged objectively; the last three are insight questions that will be judged subjectively by a retail expert. Two of the three insight questions are the same as the classification questions.

SUBMISSION INSTRUCTIONS
http://www.ecn.purdue.edu/KDDCUP/data/submissioninstructions.html

1. Given a set of page views, will the visitor view another page on the site or will the visitor leave?

Submission requirement :

The Submission for this question must be a single plain text file with the name "question1.submit". Each line in the file is for one session containing the following values separated by ",":

1. session id
2. "probability" that the visitor will view another page (number between 0 and 1, in floating point notation -- not scientific notation). The "probability" value will be thresholded at 0.5 for evaluation. Although you may submit only 0's and 1's, an

actual floating point number will help draw lift curves or other illustrative graphs.

An example submission file of this question is available at q1_submission.data
The evaluation criteria is the ratio of correct predictions on the test set (0-1 loss). **Motivation:** the ability to predict whether a visitor will leave can help determine the best page to display (e.g., there may be three alternative pages a site can display in response to a given link and each can be evaluated on the likelihood that it will cause the user to leave) or action can be taken to encourage the visitor to stay (e.g., special promotion). At the Gazelle site, no such personalization exists in the data you are being given.
2. Given a set of page views, which product brand will the visitor view in the remainder of the session?
Submission requirement :

The Submission for this question must be a single plain text file with the name "**question2.submit**". Each line in the file is for one session containing the following values separated by ",":

1. **session id**
2. **predicted brand** (string). The predicted brand (target) must be one of the following four strings: Hanes, DonnaKaran, AmericanEssentials, Other.
3. **four "probabilities"** (number between 0 and 1, in floating point notation -- not scientific notation), one for each target value in the following order: Hanes, DonnaKaran, AmericanEssentials, Other. These values will be used only for illustrative graphs and must be in the order given. **The four probabilities must match the predicted brands in the above order.**

See q2_submission.data for an example submission file of this question.
The evaluation criteria is the sum of the following units:

- 2 units for predicting Hanes when the visitor indeed visits the details page for a product of brand Hanes in the remainder of the session.
- 2 units for predicting DonnaKaran when the visitor indeed visits the details page for a product of brand DonnaKaran in the remainder of the session.
- 2 units for predicting AmericanEssentials when the visitor indeed visits the details page for a product of brand AmericanEssentials in the remainder of the session.
- 1 unit for predicting other when the visitor does not visit a details page that is of the above brands, i.e., either they view only other products or they leave prior to viewing another product.

Motivation: If you could put a single dynamic link (e.g., picture and URL) on the current page, pointing to a brand, which brand would you choose? To make the problem manageable, we are asking you to predict one of three common brands, or other.
3. Given a set of purchases over a period of time, characterize visitors who spend more than $12 (order amount) on an average order at the site.
Submission requirement: text and graphs that a business user (a retail expert will judge this) will understand and will believe is useful.

The text is limited to 1,000 words (more than one page is ok), and up to 10 graphs with captions not exceeding 3 lines each.

The evaluation criteria: how much insight does the characterization provide from a business perspective.
This is obviously a subjective evaluation. The retail expert will have access to the organizing committee members, who will be able to help interpret results (e.g., we can explain what a decision tree represents), but the ultimate decision will be left to the retail expert. **Motivation:** insight can be valuable for determining the business directions, marketing decisions, etc. **Note**: no attempt is made to hide any attributes in this dataset since the task is an insight task (not a prediction task), where we are asking for interesting insight to business users. For example, order amount (for the order) is given on every order line, which is an extremely good predictor (but not a perfect predictor if someone makes multiple purchases because the above question is for an average

purchase amount per user and multiple purchases might exist but are rare given the short time span provided). Telling a business user that the characterization of the visitors who spend more than $12 on average is that their order amount is over $12 would likely result in the termination of your consulting agreement.

4. Given a set of page views, characterize killer pages, i.e., pages after which users leave the site.
   Note: this is the insight version of question 1.

The submission requirements, evaluation criteria, and motivation are the same as question 3.

5. Given a set of page views, characterize which product brand a visitor will view in the remainder of the session?

   Note: this is the insight version of question 2.

The submission requirements, evaluation criteria, and motivation are the same as question 3.

It is important to realize that to simulate real-life situations, the training set and the test set are disjoint in time, i.e., the data provided has a cutoff date of 1 Apr 2000, and the test set begins at that date. The site has changed over time, new products were introduced, etc.

The training sets for these questions are given to you. To help you evaluate your experiments, there are two columns (question 1 testset, question 2 testset), which indicate whether those records would have been part of the test set. The two test sets that you will be given will be disjoint in time.

The test set will be release at a later time, closer to the evaluation period.

The Domain

Gazelle.com sells legware and legcare products. They went live with Blue Martini's application on January 30, 2000.

Here is what their home page looked like for the first three month. The pictures changed regularly to promote products, but the shape was the same. A major change was made early May that you can see today, but it is not relevant for the KDD Cup.

Note that the home page was very busy with about 70 images that their creative agency created.
As many dot-coms, their initial goal was to attract customers, even if it meant losing money in the short term. They had many promotions that are relevant for mining, since these effect traffic to the site, the type of customers, etc. Promotion codes that are used in orders are logged with the purchases (order header). Here are the important ones:

- FREE - free shipping ($3.95 value). Active from 3/20 to 4/30 (shipping is normally free if above $40).
- MARCH1 - $10 off from 3/1 to 4/1
- FRIEND - $10 off from 3/1 to 4/30
- FREEBAG - free bag from 3/30 to 4/30

Here are most things you should know:

- They did an Ally McBeal ad on Feb 28th. This is a prime time TV comedy show in North America and perhaps in many other countries.
- Gazelle is located in the east coast, so all time stamps are EST.
- Gazelle changed their registration form around 2/26, so some customer attributes were collected prior to this period and some after.
- If you browse the site, you will notice the "never run out" program, which automatically ships products to subscribers at regular intervals. The program began mid February and there is not enough data about it, nor is the relevant data supplied, so you can ignore it for the KDD Cup.

Data Collection and Cleaning

The Blue Martini Architecture has an application server that generates web pages from templates (.jhtml or .jsp). The architecture logs customer transactions and clickstreams, so standard sessionizing problems are not relevant (the application server assigns unique ids to sessions). All the data provided for the KDD Cup was generated by the application server logging mechanism.

Very little removal of data was done from the file. We removed the following:

- Test users were removed based on multiple criteria. We chose to do the removal because some information needed for the removal is not available in the KDD Cup dataset (e.g., user names). One criteria that was used, which we found useful now and in the past, is to flag credit card numbers that were used by more than 15 usernames and remove the users. All transactions and clickstreams by those users were also removed.
- Keynote (www.keynote.com) measures site performance. Gazelle subscribed to their service and as a result, their home page was hit about three times a minute, 24 hours a day, 7 days a week, generating about 125,000 sessions a month. We chose to remove all these sessions.
- In case of server crashes, clickstreams may not get properly logged. There are some purchases that do not have clickstream sessions, but that number is relatively small. No attempt was made to remove such purchases.
- Returned orders and uncompleted orders were removed. It's a small number and causes confusion.

Data Provided

Files are provided in extended C5.0 format with a .names file and a .data file. Some columns are marked as date, time, and ordered. You may convert these to "ignore" to run against C5.0. Note that date and time fields were split into two columns for the C5.0 format.
We also converted all commas in the data (the C5.0 separator character) to semicolons for you. This makes importing to databases easier, although you'll still have to pre-process date/time columns depending on your import tool.
There are two training set files:

- clicks - contains clickstream information. Each record is a page view. Should be used for questions 1, 2, 4, and 5.

There are two columns "Question 1 Test Set" and "Question 2 Test Set." These should help you evaluate your algorithms.
These columns contain True if you can expect that row to be in the test set when you get it from us, and False otherwise.
The clipping of sessions for the test set will be done the same wasy as it is provided in these two columns.
For question 2, each session of length n is clipped as follows:

- If the session is of length 2, then only the first request is marked as test set.
- If the session is of length n, where n is greater than 2, then a random number r is generated from 1 to n-1 and all page views with a sequence number less than or equal to r are considered to be part of the test set you would get.

For question 1, each session has a 50% chance of being clipped by the same mechanism defined above for question 2; otherwise, all the page views are marked as part of the test set you would get.
For question 1 and 4, we provided a column called "session continues" that identifies whether the session was clipped or not.
Example, suppose the session for a user consists of ten page views:

- Page 1 - home page. Assume the user typed hanes in the search on the home page
- Page 2 - search results showing hanes products Assume the user clicked on one product for details
- Page 3 - a product details page.
- .... user browses other pages, views products, etc.
- Page 10 - user views a page and then clicks on his bookmarks on yahoo.com (effectively leaving the site)

In the test set that you will get for question 1, the above set of page views has a 50% of being clipped in the middle.
If it is clipped in the middle, the clipping point will be random between page 2 to 9 (uniform distribution). If the session is not clipped, you will get all page views (1-10). If the session is clipped, you will get a subset, say page views 1 to 7. Your task is to predict whether the pages we gave you for a session are the whole session (e.g., the last page view was the last page in the session), or whether it was clipped, in which case the visitor will browse at least another page.
In the test set that you will get for question 2, every session will be clipped and you will have to determine which product details page will be viewed in the clipped pages.
The two columns provided are an example of such clipping.

orders - contain order information. Each record is an order line, which is part of an order header. Should be used for question 3. Note that clicks may also be relevant for this question, so think of joining them.

Here are descriptions of columns that may not be obvious. Many columns, especially product attributes, belong to Gazelle, not to the standard Blue Martini Schema, so we do not have a description for them, but they are usually simple to interpret. In addition to regular columns, we have enhanced the data with Acxiom attributes described below. Acxiom is a syndicated data provider and the Acxiom Data Network was used for this enhancement. To get Acxiom attributes, one must have name and address, which is usually available only about people who registered or bought something. Furthermore, Acxiom has a hit rate of about 60-70%, so few records have complete demogrpahics.

Column Name	Description
Request Processing Time	The time, in milliseconds, that it took to compile and run the .jhtml template code including all API calls. A little background should be helpful here. When the user's browser requests a page, it goes to a web server. In the Blue Martini architecture, the web servers handle all pages but .jhtml (and .jsp), which go to the application server. This means that GIFs, for example, are handled by the web server and don't bother the application server. Requests that end with .jhtml or .jsp get forwarded to the application server, which is where the logic runs. Such pages are really templates that include embedded Java code, which commonly call the Blue Martini API to determine product names, attributes, prices, promotions, etc. The templates are "executed" (they're converted to Java classes) and generate html, which is then given to the web server to serve. The "Request Processing Time" attribute measures how long it takes the application server to execute the .jhtml and convert it to .html, including the time it takes to make all the API calls. It does not include the time to actually serve the page by the web server. When a .jhtml template is hit for the first time after the application server goes up (possibly after a crash), it takes several seconds to compile the code. In addition, if the .jhtml code makes expensive API calls (e.g. search) it may take significant time to return. You may notice a bizarre runaway request of many hours in the data.
Request Query String	Query string portion of the Web page request.
Request Referrer	URL of the referring Web page.
Request Date	Date when the app server finished executing the request
Request Time	Time when the app server finished executing the request
Line Assortment ID	Unique ID of the assortment associated the lineitem (if any).
Line Subassortment ID	Unique ID of the subassortment associated withthe line item (if any).
Request Sequence	Sequence number of the request within the session. The sequence numbers begin at 1 for the first request and increase by 1for each subsequent request.
Request Template	Template page requested
REQUEST_DAY_OF_WEEK	Day of the week first request was made.
REQUEST_HOUR_OF_DAY	hour of request
Account Creation Date	The date at which the user account was created
Account Creation Date_Time	The date at which the user account was created
Truck owner to other indiv occupation	Acxiom columns, described below
Session Start Login Count	Number of times the user logged onto thewebsite at time of the session.
Session Cookie ID	Unique ID of the cookie used for the session. Has value even if customer has cookies off.
Session ID	Unique ID for the session.
Session Customer ID	Unique ID of the customer who created the session. This is set if the user either logged onto the website during the sessionor made an anonymous purchase during the session. This is NULL forusers

	who browsed anonymously without purchasing during the session.
Session User Agent	Type of browser used in the session (the user-agentstring passed from client browser).
Session Visit Count	Number of visits from the cookie at the time ofthe session.
Session First Processing Time	The length of time (in milliseconds)that it took to compile and execute the first requested template for thesession.
Session First Query String	Query string portion of the first Web pagerequest for the session.
Session First Referrer	URL of the first referring Web page for thesession.
Session First Template	First requested template in the session.
Session First Request Hour of Day	Hour of the day first request wasmade.
Product Level 1 Path	The first level of the hierarchy for the product
Product Level 2 Path	The first two levels of the hierarchy for theproduct
Product Level 3 Path	The first three levels of the hierarchy for theproduct
Assortment columns	Describing the assortment
Content	Attributes about the content pages
Viewed Brand	For every page view this is set to AmericanEssentials,DonnaKaran, Hanes, Other, or "?"

The following are the Acxiom columns

Truck Owner	True if the customer owns a truck. Many of these specialtyvehicle owners display a greater-than-average interest in outdoors anddo-it-yourself activities.
RV Owner	True if the customer owns a recreational vehicle.
Motorcycle Owner	True if the customer owns a motorcycle.
Value Of All Vehicles	Value of all the vehicles owned by the customer(based on Blue Book value of all vehicles registered in the household).
Age	Age of the customer.
Other Indiv. Age	Age of the other individual in the customer's household(if any).
Marital Status	Marital status of the customer.
Working Woman	True if there is a working woman in the household.
Mail Responder	True if the customer has responded to any direct marketingmail campaigns.
Bank Card Holder	True if the customer has a bank card (e.g. VISA orMastercard).
Gas Card Holder	True if the customer has a gas card or a retail storecard.
Upscale Card Holder	True if the customer has a credit card from anupscale retail store.
Unknown Card Type	True if the customer has a credit card of unknowntype.
TE Card Holder	True if the customer has a travel and entertainmentcard.
Premium Card Holder	True if the customer has a premium credit card.(such as Gold or Platinum card with a high credit line).
Presence Of Children	True if there are children present in the household.
Number Of Adults	Number of adults in the customer's household.
Estimated Income Code	Estimated income range for the customer's household.
Home Market Value	Estimated range of the value of the home.
New Car Buyer	True if the customer is a new car buyer.
Vehicle Lifestyle	The dominant lifestyle for the customer's vehicle(s).
Property Type	The type of property the customer owns.
Loan To Value Percent	LTV ratio is the percentage of outstanding homeloan balance in relation to home's market value. Based on modeling of currentand historical mortgages compared to market value.
Presence Of Pool	True if the customer has a pool at this address.
Year House Was Built	Year that the customer's house was built.
Own Or Rent Home	Indicates if the customer is an owner or renter.
Length Of Residence	Length of time that the customer has stayed atthis address.
Mail Order Buyer	True if the customer has bought something by mailorder.
Year Home Was Bought	The year that the customer's house was bought.
Home Purchase Date	The year and month in which the customer's housewas bought
Number Of Vehicles	Number of vehicles in the customer's household.

DMA No Mail Solicitation Flag	True if the customer does not want toreceive mail.
DMA No Phone Solicitation Flag	True if the customer does not wantto receive phone calls.
CRA Income Classification	1: LOW INCOME 2: MODERATE INCOME 3: MIDDLE INCOME 4: HIGHER INCOME
New Bank Card	True if the customer has a new credit card.
Number Of Credit Lines	Number of credit lines that the customer has.
Speciality Store Retail	True if the customer had any retail activityat a specialty store recently.
Oil Retail Activity	True if the customer had any oil/gas retail activityrecently.
Bank Retail Activity	True if the customer had any bank retail activityrecently.
Finance Retail Activity	True if the customer had any finance retailactivity recently.
Miscellaneous Retail Activity	True if the customer had any misc. retailactivity recently.
Upscale Retail	True if the customer had upscale retail activity recently.
Upscale Speciality Retail	True if the customer had upscale specialtyretail activity recently.
Retail Activity	True if the customer had any retail activity recently.
Last Retail Date	Most recent date of retail activity.
Dwelling Size	Single or Multi-Family dwelling.
Dataquick Market Code	Relative home market value indicator that compare sthe home with others in the same county (Top 10% ,Top 20%, etc.).
Lendable Home Equity	Amount of home equity that can be used for aloan. LHA is 80% of a home's market value less total amount of loans.
Home Size Range	Range of the size of the customer's house.
Lot Size Range	Range of the size of the customer's lot.
Insurance Expiry Month	Month in which the customer's homeowner insuranceexpires.
Dwelling Unit Size	Counts the number of known households at the customer'saddress.
Month Home Was Bought	Month in which the customer bought the house.
Household Status	Indicates if the customer is the first or secondindividual in the household.
Verification Date	Last time that this customer's Acxiom data was verified."
Minority Census Tract	
Year Of Structure	The year that the first structure was built at thecustomer's address.
Gender	Customer's gender.
Occupation	Customer's occupation.
Other Indiv. Gender	Gender of the other individual in the customer'shousehold (if any).
Other Indiv. Occupation	Occupation of the other individual in thecustomer's household.

Here are line item related attributes:

Order Line Session ID	Unique ID of the session in which the orderwas placed. This is NULL if clickstream for the session was not logged.
Order Line Date	Date the order line was created
Order Line Unit List Price	The list price amount for one unit of anorder line item
Order Line Assortment ID	Unique ID of the assortment through whichthe item was purchased. This is NULL if if the item was not

	purchasedthrough an assortment.
Order Line Subassortment ID	Unique ID of the subassortment throughwhich the item was purchased. This is NULL if the item was not purchasedthrough a subassortment.
Order Line ID	Unique ID of the order line.
Order Line Quantity	Quantity of units purchased in the order line
Order Line Unit Sale Price	Sale price amount for one unit of the lineitem
Order Line Status	Order line status
Order Line Tax Amount	Total tax amount for the order line
Order Line Amount	Total amount for the order line (tax amount + quantity* unit sale price).
Order Line Day of Week	Day of the week order line was placed
Order Line Hour of Day	Hour of the day order line was placed
WhichDoYouWearMostFrequent : Description" String "City from customeraddress	
Account Creation Date	Date customer account was created
Email	Customer's e-mail address
Login Failure Count	Number of failed logins for account
Account Status	Status of the customer account
Customer ID	Unique ID for the customer account
Order Date	Date the order was placed.
Order Customer ID	Unique ID of the customer who placed the order
Order Discount Amount	Discount amount for the order
Order Modification Date	Date the order was last modified
Order System Number	Unique order system number.
Order ID	Unique ID of order header
Order Session ID	Unique ID of the session in which the order line wasplaced. This is NULL if clickstream for the session was not logged
Order Promotion Code	Optional promotion code to provide additionalorder discounts
Order Promotion ID	Unique ID of the promotion associated with the order. This is NULL if none.
Order Source	Source through which the order was placed
Order Status	Order status
Order Amount	Total amount for the order (discount amount + shippingamount + tax amount).
Order Shipping Amount	Total shipping amount for the order
Order Tax Amount	Total tax amount for the order
Order Shipment Method ID	Unique ID for shipment method
Order Day of Week	Day of the week order was placed
Order Hour of Day	Hour of the day order was placed

Change Log

This section describes changes to this file, so that if you have read it once, you can see the updates.

- 5/21/00 - Question 3 was clarified to say "$12 (order amount) on an average order at the site."

The following note was also added for this question:

> **Note**: no attempt is made to hide any attributes in this dataset since the task is an insight task (not a prediction task), where we are asking for interesting insight to business users. For example, order amount (for the order) is given on every order line, which is an extremely good predictor (but not a perfect predictor if someone makes multiple purchases because the above question is for an average purchase amount per user and multiple purchases might exist but are rare given the short time span provided). Telling a business user that the characterization of the visitors who spend more than $12 on average is that their order amount is over $12 would likely result in the termination of your consulting agreement.

- 5/21/00 - The never-run-out program was clarified as follows:
  If you browse the site, you will notice the "never run out" program, which automatically ships products to subscribers at regular intervals. The program began mid February and there is not enough data about it, nor is the relevant data supplied, so you can ignore it for the KDD Cup.
- 5/21/00 - An explanation was made Ally McBeal is a prime time TV comedy show in North America and perhaps in many other countries.
- 5/21/00 - The following sentence was added to the questions section:
  The two test sets that you will be given will be disjoint in time.
- 5/21/00 - An example was given in the data section.
- 5/26/00 - Explained that session continues is the target for question 1 and 4.
- 5/27/00 - Added a long explanation for request processing time following a question
- 5/30/00 - Added an explanation for the Acxiom column "CRA Income Classification".
- 7/13/00 - Added submission example files and clarified the probability ordering requirement for questions 1 and question 2, which require scoring.
- 7/13/00 - Add file format requirement for submissions to question 1 and question 2.
- 7/13/00 - Add submission instructions link:
http://www.ecn.purdue.edu/KDDCUP/data/submissioninstructions.html
- 7/14/00 - We loosened the 1-page restriction on submissions to questions 3, 4, and 5. You can have more than one page for your submission to each of these 3 questions, but the 1,000 word limit is still required.

APPENDIX C: SAMPLE SQL QUERIES

Table C. 1 A stored procedure code segment which updates the column names of the clicks table

```
EXEC sp_rename 'KDD2000CLICKS.[col001]', 'RequestProcessingTime', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col002]', 'RequestQueryString', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col003]', 'RequestReferrer', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col004]', 'RequestDate', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col005]', 'RequestDate_Time', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col006]', 'RequestAssortmentID', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col007]', 'RequestSubassortmentID', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col008]', 'RequestSequence', 'COLUMN'
.............
.............
EXEC sp_rename 'KDD2000CLICKS.[col212]', 'ContentLevel2Path', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col213]', 'ContentLevel3Path', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col214]', 'Question1TestSet', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col215]', 'Question2TestSet', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col216]', 'SessionContinues', 'COLUMN'
EXEC sp_rename 'KDD2000CLICKS.[col217]', 'ViewedBrand', 'COLUMN'
```

Table C. 2 Aggregation of derivate variables from clicks table

```
SELECT
MAX(cast(dbo.clicks.RequestSequence as int)) AS clicks,
MIN(dbo.clicks.RequestDate_Time) AS begtime,
MAX(dbo.clicks.RequestDate_Time) AS endtime,
AVG(CAST(dbo.clicks.RequestProcessingTime AS int)) AS processtime,
dbo.clicks.SessionID
into clicksaggregate

FROM        dbo.clicks INNER JOIN
                 dbo.orders ON dbo.clicks.SessionID = dbo.orders.OrderSessionID
GROUP BY dbo.clicks.SessionID
order by sessionID
```

Table C.3 Calculation of derivate variables from clicks table

```
update clicksaggregate
set duration=datepart( minute,(endtime-begtime))+datepart( hour,(endtime-begtime))*60
```

Table C. 4 Update operation of the fact table by derivate variables.

```
update ordersfacttable
set
duration=clicksaggregate.duration,
clicks=clicksaggregate.clicks,
processtime=clicksaggregate.processtime
from clicksaggregate
where
ordersessionID=sessionID
```

Table C. 5 A query that replaces missing cases with valid null.

```
update dbo.anorderscustomerdimension
set
ValueOfAllVehicles=null
where ValueOfAllVehicles=0
```

Table C. 6 A query that removes excessive characters from the variable.

```
update orderstest
set
AccountCreationDate=replace(AccountCreationDate,'\',':'),
AccountCreationDate_Time=replace(AccountCreationDate_Time,'\',':'),
AccountStatus=replace(AccountStatus,'\',':'),
ActionCode=replace(ActionCode,'\',':'),
AcxiomID=replace(AcxiomID,'\',':'),
Age=replace(Age,'\',':'),
.....
```

Table C.7 A query that investigates the relation between fact table and dimension in the given conditions.

```
select count(*)as nullquesitons from a1factable
where
(customerID in
(select customerID from dbo.a1customers
where YourFavoriteLegcareBrand='' and HowOftenDoYouPurchase=''))
```

Table C.8 A query that investigates the relation between fact table and dimension in the given conditions.

```
SELECT     COUNT(*) AS facttablevalidation
FROM       dbo.anordersproductsdimension
WHERE     (ProductID IN
                  (SELECT     productID
                    FROM        dbo.anordersfacttable))
```

Table C.9 A query that creates a new table.

```
SELECT          ProductHierarchyLevel, UnitsPerInnerBox, PrimaryPackage, StockType, Pack,
BasicOrFashion, LeadTime, HasDressingRoom, Manufacturer, Material,
                Collection, ProductWebTemplate, Audience, ProductFamilyStatus, ProductID,
ProductStatus, ProductLevel1Path, ProductLevel2Path,
                ProductLevel3Path, AssortmentType, AssortmentID, AssortmentStatus,
AssortmentModificationDate, AssortmentLevel, VendorPageColor,
                AssortmentLevel2Path, AssortmentLevel1Path, SpendOver$12PerOrderOnAverage,
AssortmentCreationDate, UnitsPerOuterBox, BrandName,
                OrderLineSessionID, OrderLineID
into ordersproducts
FROM       dbo.orderstest
```

Table C.20 A query that updates a table's column from another table's column.

```
update orderstest
set
orderstest.orderlinedate_time=orders.orderlinedate_time,
orderstest.orderlinedate=orders.orderlinedate
from orders where orderstest.keyorder=orders.keyorder
```

APPENDIX D: SPSS ANSWER TREE SOLUTION SUMMARY

Project Information
Project File
F:\kdd2000data\answertree\iyiagac.atp

Name of Tree
SPENDOVER12

Data File Server
Local Computer

Server Specification
GET
  FILE='F:\kdd2000data\answertree\customer2_2.sav'.

Number of Cases

	Weighted	Unweighted
Cases	1354,00	1354,00

Partition Information
Partition
Off

Cross Validation Information
Cross-Validation
Off

Tree Growing Criteria
Growing Method
Exhaustive_CHAID

Algorithm Specifications

Alpha for splitting:	0,05
Chi-squared statistic:	Pearson
Allow splitting of merged criteria:	Off
Use of Bonferroni adjustment:	On

Stopping Rules

Maximum tree depth:	7
Minimum no. of cases for parent node:	40
Minimum no. of cases for child nodes:	20

Model
Target Variable

Name	SPENDOVE
Label	spendover12perorder
Type	String

Measurement Level Nominal

Predictors

Name	Type	Level	Label
YOURFAVO	String	Nominal	YourFavoriteLegcareBrand
WHICHDOY	String	Nominal	WhichDoYouWearMostFrequent
HOWDOYOU	String	Nominal	HowDoYouDressForWork
YEARHOME	Numeric	Continuous	YearHomeWasBought
NUMBER1	Numeric	Continuous	NumberOfVehicles
GENDER	String	Nominal	Gender
OTHERI1	String	Nominal	OtherIndiv_Gender
MARITALS	String	Nominal	MaritalStatus
HOWMNYP	Numeric	Ordinal	HowManyPairsDoYouPurchase
SENEMA	Numeric	Nominal	SendEmail
HOWOFDYP	Numeric	Ordinal	HowOftenDoYouPurchase
HOMEMARV	Numeric	Ordinal	HomeMarketValue
ESTINC4	Numeric	Ordinal	estimated income

Cost
True
False
*

True
0,0000 1,0000
False
1,0000 0,0000
Profits
True
False
*

Revenue	0,0000	1,0000
Expense	0,0000	0,0000
Profit	0,0000	1,0000

*Target Level

Resulting Tree
Size

Total number of nodes	18
Total number of levels	4
Total number of terminal nodes	11

CURRICULUM VITAE

Osman Onat Ünal was born in 1976, Istanbul. He graduated from Beşiktaş Atatürk Anadolu High School in 1994. He became an engineer after taking Bachelor of Science Degree from Istanbul Technical University Management Engineering Department in 1999.

He worked as a software project manager in the private sector during 1999-2000. After this experience he worked as a research assistant in Boğaziçi University Management Information Department during 2000-2003. At the same time, he began his graduate education in Istanbul Technical University Institute of Science and Technology Management Engineering Programme in 2000 and graduated in 2003.