NECESITATEA UTILIZĂRII TEHNICILOR DATA MINING ÎN E-BUSINESS

Claudia Elena Dinucă
Facultatea de Economie şi Administrarea Afacerilor,
Universitatea din Craiova
clauely4u@yahoo.com

Rezumat :

Numărul utilizatorilor de Internet a crescut de la 400 de milioane în 2000 la puţin peste 2 miliarde la începutul lui 2011. Aceasta înseamnă că aproximativ o treime din populaţia globului utilizează Internetul. În aceste condiţii modul în care sunt gândite afacerile trebuie schimbat.

Multe firme, care, în condiţiile secolului trecut, nici măcar nu puteau visa că ar putea să aibă un anumit volum de activitate sau că ar putea să facă faţă concurenţei giganţilor din industria lor, au reuşit să se bucure de un mare succes. Putem da doar un exemplu: Amazon.com, înfiinţată în 1995, avea în 1999 o cifră de afaceri de cel puţin 13 ori mai mare decât alte nume de prestigiu din SUA, cum ar fi Barnes & Nobles şi Borders Books & Music.

E-business este cheia pentru a face viaţa mai uşoară pentru oameni.

Cunoaşterea mediului e-business este esenţială pentru a putea face afaceri în acest secol. Mai mult, trebuie înţelese şi aplicate noile tehnologii de extragere a cunoştinţelor din date.

***CUVINTE CHEIE**: data mining, clusterizare, regresie, reguli de asociere, e-business.*

1. Introducere

Istoric, noţiunii determinării modelelor (informaţii inteligibile) din date i-a fost data o varietate de nume de către statisticieni şi comunitatea profesională care lucrează cu baze de date şi exploatări de date (data mining), cunoştinţe de data mining, descoperirea de informaţii, recoltararea informaţiilor, arheologia şi prelucrarea datelor, forme (modele) de date. Sistemul de descoperire a cunoştinţelor, care este capabil să lucreze în sistemul de baze de date de mari dimensiuni se numeste Descoperirea Cunoştinţelor în sisteme de baze de date -KDD.

Termenul KDD a apărut pentru prima dată în 1989. Prin definiţie, KDD este "un proces non-triviale de extragere a informaţiilor, date anterior necunoscute şi potenţial utile" dar şi ca "ştiinţa de a

THE NEED TO USE DATA MINING TECHNIQUES IN E-BUSINESS

Claudia Elena Dinucă
Faculty of Economics and Business Administration,
University of Craiova
clauely4u@yahoo.com

Abstract :

The number of Internet users rose from 400 million in 2000 to just over 2 billion in early 2011. This means that approximately one third of the world's population uses the internet. Taking these conditions into consideration, the way how businesses are designed need to be changed

Many companies that, over the last century could not even dream that could have a certain volume of activity or they could face competition with industry giants, have succeeded in giving to enjoy great success. For example: Amazon.com, founded in 1995, had in 1999 a turnover of at least 13 times higher than other prestigious names in the U.S., such as Barnes & Noble and Borders Books & Music.

E-business is the key to make life easier for the people.

Knowledge of e-business environment is essential for doing business in this century. More must be understood and new technologies applied to extract knowledge from data.

***KEYWORDS**: data mining, clustering, regresion, asociation rule, e-business.*

1. Introduction

Historically, the notion of determining patterns (understandable information) of data was given a variety of names by statisticians and community professionals working with databases and data mining (data mining), knowledge of data mining, discovery information, harvesting information, data archeology and processing forms (patterns) of data. The Knowledge Discovery System which is able to operate on large scale database system is called Knowledge Discovery in Databases System-KDD.

KDD term first appeared in 1989. By definition, KDD is „a non-trivial process of extracting information, previously unknown and potentially useful data" but as „the science of extracting

extrage informaţii utile din masive de baze de date", dupa Fayyad şi alţii, 1996. În acest context, datele sunt o colecţie de fapte, şi modelul este un nivel superior de exprimare care descrie datele sau un subset din acestea. În analiza datelor, caracteristicile modelelor pe care KDD le identifică trebuie să fie valide, de noutate, neredundante, folositoare şi în cele din urmă de înţeles. Un model corect este acela care descrie datele cu ceva grade de siguranţă. În final, este de dorit ca modelele descoperite să fie de înţeles pentru a putea fi analizate ulterior pentru a studia cauzele şi efectele. Deoarece extragerea de date (Data Mining) este partea centrală a procesului de descoperire de cunoştinţe din bazele de date (KDD), termenii data mining şi descoperirea de cunoştinţe din baze de date au fost utilizaţi alternativ de mulţi cercetători din domeniu. În ultimul timp însă, se face o distincţie clară între cei doi termeni. Distincţia care se face este referitoare la faptul că descoperirea de cunoştinţe din bazele de date (KDD) poate fi considerată ca procesul de extragere a informaţiilor folositoare şi interesante din baza de date. Autorii care fac deosebire între DM şi KDD consideră KDD ca fiind un process iterativ şi interactiv complex, care include DM. KDD se referă la procesul de descoperire a cunoştinţelor folositoare din date, în timp ce data mining se referă la un pas particular din acest proces. Data mining reprezintă aplicarea unor algoritmi specifici pentru extragerea pattern-urilor (modelelor) din date.

Drept o consecinţă a disponibilizării marilor rezervoare de date s-a dezvoltat data mining. Colectarea datelor în diverse formate de digitizare a început în anii '60 permiţând o analiză retrospectivă a datelor prin intermediul calculatorului. În anii '80 au aparut bazele de date relaţionale împreuna cu Structured Query Language (SQL) permiţând analizarea dinamica la cerere a datelor. Anii '90 sunt caracterizaţi de o explozie a datelor. Pentru stocarea lor au început să se folosească depozitele de date (data warehouses). Drept răspuns la provocările cu care s-a confruntat comunitatea specialiştilor în baze de date a apărut data mining, care se ocupă cu cantităţi masive de date, aplicarea analizei statistice şi aplicarea tehnicilor de cautare specifice inteligenţei artificiale asupra datelor. Rolul data mining este extragerea de cunoştinţe noi, implicite şi cu acţiune directă din colecţii mari de date, descoperirea lucrurilor care nu sunt evidente din date, care nu pot fi extrase manual, reprezentând informaţii folositoare care pot îmbunătăţi procesul curent de acţiune.

useful information from massive data or databases" according to Fayyad and others, 1996. In this context, the data is a collection of facts, and the model is a higher level of expression that describes the data or a subset there of. The data analysis features of models that KDD identifies must be valid, novelty, without repetitions**,** useful and ultimately understandable. A model is correctly describing the data with some degree of safety. Finally, it is desirable that the models found to be understood so being further analyzed to study the causes and effects. Because Data Mining is the central part of the process of knowledge discovery from databases (KDD), the terms data mining and knowledge discovery in databases were used alternately for many researchers in the field. Lately, however, is a clear distinction between the two terms. The distinction is related to that of knowledge discovery in databases (KDD) can be considered as the extraction of useful and interesting information from the database. The authors who distinguish between DM and KDD as KDD is considered an iterative and interactive complex process that includes DM. KDD refers to the process of discovering useful knowledge from data, while data mining refers to a particular step in this process. Data mining represents the application of specific algorithms for extracting patterns (models) of data.

As a consequence of the dismissal of large reservoirs of data has developed data mining. Collecting data in various formats, digitization began in the 60s allowing a retrospective analysis of data by computer. In the 80s came relational databases with Structured Query Language (SQL) and application that allows dynamic data analysis. The 90s years are characterized by an explosion of data. To store them it began to use data warehouses. In response to the challenges faced by the community of specialists in database data mining appeared, dealing with massive amounts of data, applying statistical analysis and search techniques specific to artificial intelligence on the data. The role of data mining is the extraction of new knowledge, implicit and direct action of large data collections, discovering things that are not obvious from the data, which can not be extracted manually, representing useful information that can improve the current action process.

2. Procesul Descoperirii Cunoştinţelor în Baze de Date

În concordanţa cu Fayyad şi colaboratorii(1996), KDD este procesul de folosire a bazei de date împreună cu etapele cerute de selectare, pre-procesare, transformare a datelor pentru a aplica metode data-mining, respectiv algoritmi, pentru a obţine pattern-uri din date şi a evalua procesul de data mining pentru a identifica subsetul de pattern-uri enumerate considerate „cunoştinţe". Procesul de KDD este divizat în şapte paşi dupa cum urmează:

1. *Analiza Domeniului* - natura datelor din domeniu este analizată şi se defineşte ţinta descoperirii. Dacă există cunoştinţe anterioare în acest domeniu, aceste sunt evaluate.
2. *Selectarea* - selectarea sau segmentarea datelor în concordaţă cu anumite criterii, ceea ce poate înseamna eliminarea unor câmpuri sau rânduri din date sau ambele.
3. *Preprocesarea* - stadiul de curăţare a datelor în care anumite informaţii sunt îndepărtate, de asemenea se determină metode de lucru cu câmpuri de date lipsă.
4. *Transformarea* - Datele sunt transformate. O reprezentare a datelor care este compatibilă cu algoritmul data-mining ce urmează a se implementa se realizează în această etapă. Datele sunt analizate cu scopul determinării unor caracteristici pentru a reprezenta datele în concordanţă cu ţinta ce trebuie atinsă.
5. *Data mining* - această etapă se ocupă de extragera modelelor din date. În acest scop se vor utiliza algoritmi data-mining adecvaţi. Calitatea acestei etape depinde foarte mult de etapele precedente.
6. *Interpretarea şi evaluarea* - Modelele identificate de sistem, în urma algoritmului aplicat, sunt interpretate în cunoştinţe care pot fi folosite să suporte deciziile luate de om, de exemplu predicţiile şi problemele legate de clasificare, sumarizând conţinutul bazei de date şi explicând fenomenele observate.
7. *Consolidarea Cunoştinţelor Descoperite* – modelele descoperite sunt puse în folosinţă. Un mod de utilizare plauzibil îl reprezintă încorporarea cunoştinţelor obţinute într-un alt sistem pentru acţiuni suplimentare, documentarea modeleleor şi transmiterea lor părţilor interesate, precum şi reaplicarea KDD bazelor folosind drept fundament aceste noi cunoştinţe.

Data mining se concretizează aşadar prin aplicarea unor algoritmi specifici pentru extragerea modelelor din date. Paşii suplimentari ai procesului de Descoperire a Cunoştinţelor din Date, cum sunt prepararea datelor, selectarea datelor, etapa de

2. Knowledge discovery in database process

According to Fayyad and his colleagues (1996), KDD is the process of using database along with the steps required as select, pre-processing, transformation of data to apply data-mining methods (algorithms) in order to obtain patterns of data and evaluate data mining process to identify the subset of patterns listed as knowledge. KDD process is divided into seven steps as follows:

1. *Domain analysis* - the nature of field data are analyzed and defined target discovery. If they are previous knowledge in this area, these are evaluated.
2. *Selection* - or segmentation of data in accordance with certain criteria, which may mean removing some fields or rows of data, or both.

3. *Preprocessing* - data cleansing stage where certain information is removed, also determine ways of working with missing data fields.

4. *Transformation* - the data is processed. A representation of the data that is compatible with data-mining algorithm that is to be implemented is done at this stage. The data is analyzed to determine the characteristics to represent data in accordance with the target to be reached.

5. *Data mining* - this step takes care of extracting the data model. For this purpose we use a data-mining algorithm properly. The quality of this phase depends heavily on the previous stages.

6. *Interpretation and evaluation* - identified system models, following the algorithms applied, are interpreted in the knowledge that can be used to support decisions made by humans, such predictions and classification problems, summarizing database content and explaining observed phenomena.
7. *Enhancing Knowledge Discovered* - models (patterns) found are put in use. A plausible way to use is the incorporation of knowledge obtained in another system for further action, documentation and transmission of models to stakeholders and reapply KDD database using this new knowledge as a basis.

Data mining is thus materialized by applying algorithms to extract patterns from data. Additional steps of the process of discovering knowledge from data such as data preparation, data selection, cleaning phase, the integration of previous knowledge required are essential to ensure that will extract useful knowledge from data.

curățare, integrarea cunoştinţelor anterioare necesare sunt esențiali pentru a asigura că se vor extrage cunoştinţe folositoare din date.

3. Tehnici data mining

Există două clase fundamentale de metode de învățare:

- *predictive* (bazate pe *învățare supervizată*), ce utilizează un set de variabile (numite predictori) prin intermediul cărora se realizează predicții relative la valorile (continue sau discrete) ale altor variabile (numite variabile de decizie);

-*descriptive* (bazate pe *învățare nesupervizată*), destinate extragerii unor patternuri (structuri inteligibile) din date.

Modelele predictive bazate pe inteligență artificială se construiesc în cadrul unei faze de antrenare, prin care modelul învață să prezică răspunsul potrivit (decizia), când la intrare se prezintă diverse seturi de valori ale predictorilor. După consumarea fazei de antrenare, modelul poate fi folosit în predicție, pentru a rezolva, după caz, fie probleme de clasificare (dacă variabila de decizie este nominală sau discretă), fie probleme de regresie (dacă variabila de decizie este continuă).

Metodele data mining descriptive formează a doua mare categorie din data mining. Spre deosebire de modelele predictive, metodele descriptive (precum cele de clustering) tratează uniform variabilele, fară să distingă între predictori şi răspuns (decizie), ca atare învățarea nu este supervizată (în sensul învățării din exemple, adică al furnizării de răspunsuri în cadrul fazei de training). Metodele descriptive permit descrierea şi explicarea fenomenelor caracteristice sistemului studiat pe baza modelelor descoperite

Reguli de asociere

Mulțimile frecvente de articole/link-uri pot fi determinate dacă luăm în considerare principiul cheie al *monocității(monocity)* sau *a-priori* care spune că dacă o mulțime de articole/item-uri(link-uri) L este frecventă (apare cel puțin în a l-a parte a site-ului/click-ului), atunci orice submulțime este tot frecventă.

Se utilizează termenul *mulțimi frecvente de articole(frequent itemsets)* pentru „o mulțime de articole S care apare în cel putin a „s"-a parte din coşuri, unde s este o constantă aleasa, de obicei 0.01.

Pentru a determina mulțimile frecvente de articole/link-uri se parcurg etapele:

1. Procedăm nivel cu nivel, găsim întâi articole/link-urile frecvente, mulțimi de dimensiune 1, apoi perechi frecvente, triplete frecvente, etc.
2. Găsim toate *mulțimile frecvente de articole/link-uri maximale* (mulțimile M astfel încât o mulțime care include strict pe M nu este frecventă) într-o singură trecere sau mai multe.

3. Data mining techniques

There are two fundamental classes of learning methods:

 predictive (based on supervised learning), which uses a set of variables (called predictors) through which predictions are made relative to the values (continuous or discrete) of other variables (called decision variables);

 descriptive (based on unsupervised learning), for extraction of patterns (structures understandable) of data.

Predictive models are built based on artificial intelligence in a training phase, in which the model learns to predict the right answer (decision) when the input values is formed with different sets of predictors. After consuming training phase, prediction model can be used to solve, as applicable to classification problems (if the decision variable is nominal or discrete) or regression problems (if the decision variable is continuous).

Descriptive data mining methods form the second largest category of data mining. Unlike predictive models, in descriptive methods (such as clustering) the variables are treated uniformly, without distinguishing between predictors and response (decision) as such is not supervised learning (in terms of learning from examples, that of providing responses in the training phase). Descriptive methods allow the description and explanation of the characteristic phenomena of the system studied based on the patterns found.

Association Rules

Crowds frequent articles / links can be determined if we consider the key principle of monocity or a priority which says that if a set of items (links) L is frequent (at least in the l-part of the site/click), then any subset is frequent. It uses the term frequent sets of items (frequent itemsets) for a set of articles S appearing in at least **s** part of the shopping basket/links, where **s** is a chosen constant, usually 0.01.

To determine frequent sets of articles/links must go through stages:

1.Proceed at the level we find the first articles/links frequent sets of size 1, then frequent pairs, triplets common, and so on.

2. Find all frequent sets of maximum articles / links (sets M so that any set strictly including M is not frequent) in one or more pass.

Metoda se poate aplica în orice sector de activitate pentru care este necesară găsirea unor grupări posibile de produse sau servicii: servicii bancare, servicii de telecomunicaţii. Poate fi aplicată în domeniul medical pentru studiul complicaţiilor apărute datorită asocierii unor medicamente sau în domeniul fraudelor, caz în care se caută asocieri neobişnuite.

Regulile de asociere se definesc astfel:

Fie I={i1,i2,...,im} un număr de simboluri, numite **elemente**. Se consideră D o mulţime de tranzacţii, în care fiecare tranzacţie T se constituie ca o submulţime al lui I, unde T este o mulţime inclusă sau egală cu I. Se iau în considerare doar prezenţa (reprezentată binar) a elementelor în tranzacţie şi nu se consideră alte caracteristici cantitative sau calitative ale acestora. Fiecărei tranzacţii îi este asociat un identificator (tid).

Măsurile cheie în cadrul extragerii regulilor de asociere sunt suportul şi încrederea. Suportul se referă la proporţia în care o relaţie apare în date. Confidenţa / încrederea regulilor de asociere se referă la probabilitatea de a găsi un antecedent având o consecinţă.

Determinarea regulilor de asociere se face în doi paşi:

- Determinarea seturilor de elemente frecvente, cele care au suport suficient;
- Determinarea regulilor de asociere dintre aceste seturi, determinarea de reguli tari . Acest pas se rezolvă astfel: pentru fiecare set frecvent X şi pentru fiecare subset al lui X ,$Y \subset X$ se determină parametrii regulii $X \setminus Y \rightarrow Y$ tinând cont ca rezultatul reuniunii părţii stângi cu partea dreapta trebuie să reprezinte un set frecvent, în acest caz $X \setminus Y \cup Y = X$.
- Regulile de asociere se folosesc pentru a găsi mulţimile frecvente de articole în bazele de date ce conţin tranzacţiile consumatorului, problemă cunoscută sub denumirea de analiza coşului de cumpărături(market basket analysis). Analiza coşului de cumpărături constă în găsirea de asocieri între produsele cumpărate, respectiv afişate pe bonul de casa. Se studiaza astfel ce cumpărături fac clienţii pentru a obţine informaţii asupra produselor ce tind a fi cumpărate în acelaşi timp. În acest caz, baza de date cu tranzacţiile consumatorilor este reprezentată printr-o secvenţă de tranzacţii T=(t1,...tn), iar fiecare tranzacţie este o mulţime de articole. De exemplu, în cazul coşului de cumpărături se poate cere ca încrederea să fie semnificativ mai mare decât în cazul în care articolele ar fi plasate aleator în coş. Se poate găsi o regulă {lapte, unt} => paine pe principiul că multă lume cumpără pâine, însa exemplul bere/scutece descoperit în SUA arată că regula {scutece}=>{bere} este verificată cu o încredere

The method can be applied in any sector which requires finding the possible groups of products or services: banking, telecommunications services. It can be applied to study medical complications due to the combination of drugs or fraud, in which case looks for unusual combinations.

Association rules are defined as follows.

Let I = {i1, i2 ,..., im} be a set of symbols, called **elements**. D is considered a set of transactions where each transaction T is a subset of I. Consider only the present (represented binary) elements in the transaction and does not consider other quantitative or qualitative characteristics thereof. Each transaction gets an identifier (tid).

Key measures in the mining association rules are support and confidence. Support refers to the proportion in which a relationship occurs in data. The confidente/trust of the association rule relates to the probability of finding an antecedent having a consequence.

Determination of association rules is done in two steps:

- Determination of frequent sets of elements, those that have sufficient support;
- Determination of rules of association between these sets of rules determining the strong rules. This step resolves as follows: for each frequent set X and every subset of X, $Y \subset X$ determine the parameters of the rule $X \setminus Y \rightarrow Y$ considering the outcome of the meeting left with the right side must be a frequent set, in this case $X \setminus Y \cup Y = X$.
- Association rules are used to find frequent sets of articles in databases that contain consumer transactions, the problem known as the market basket analysis. Market basket analysis consists of finding associations between items purchased, displayed on the receipt. It studies how the customers are doing shopping to get information on the products which tend to be purchased at the same time. In this case, the database of consumer transactions is represented by a sequence of transactions T = (t1, ... tn), and each transaction is a set of articles. For example, in the case of shopping cart it requires that trust to be significantly higher than if the items were placed randomly in cart. It can be found a rule {milk, butter} => bread on the principle that many people buy bread, but the example of beer / diaper found in the U.S. show that the rule {diapers} => {beer} is checked with a significantly higher confidence than multitude of baskets containing beer. The result of this study helps retailers in the settlement of the articles in shelves and controls how a typical buyer crosses store

semnificativ mai mare decât a mulţimii de coşuri conţinând bere. Rezultatul acestui studiu ajută vânzătorii în procesul de aşezare a articolelor în rafturi şi controlează modul în care un cumpărator tipic traversează magazinul.

În cazul analizei click-urilor se lucrează pe o bază de date cu sesiunile serverului care înregistrează solicitările utilizatorilor. Sesiunile utilizatorilor sunt secvenţe S=(s1,...,sn) formate cu paginile vizitate de utilizator. Determinarea link-urilor frecvente şi a regulilor de asociere este esenţială pentru problema analizei click-urilor, modul în care utilizatorii navighează pe Internet şi accesează diverse site-uri.

Reguli de asociere secvenţiale

De multe ori, tranzacţiile sunt înregistrate ţinând cont de o secvenţă temporală. De exemplu, tranzacţiile pentru deţinătorii unui card de loialitate corespund secvenţei de chitanţe de vânzare. Tranzacţiile care înregistrează căile de navigare urmate de către un anumit utilizator web sunt asociate cu o secvenţă temporală a sesiunilor. În astfel de situaţii, analiştii sunt interesaţi să extragă regulile de asociere care iau în considerare dependenţe temporale. Problema descoperirii regulilor secvenţiale a fost introdusă prima dată de către Agrawal şi Srikant în [2].

Clasificarea şi regresia sunt fome de învăţare supervizată. Clasificarea şi regresia reprezintă cea mai largă categorie de aplicaţii, constând în construirea de modele în scopul previzionării apartenenţei la un set de clase (clasificare) sau a unor valori (regresie). Există câteva tehnici dedicate rezolvării problemelor de clasificare şi regresie, dintre care arborii decizionali, tehnica Bayes, reţelele neuronale şi k-NN se bucură de o largă recunoaştere.

Tehnicile de învăţare supervizate au drept scop să genereze mecanisme de inducţie automată cu mare putere predictivă prin extragerea informaţiilor conţinute în baza de date şi transformarea lor într-o baza de cunoştinţe. Există două mari clase de algoritmi de inducţie :

 Algoritmi de clasificare - când variabila în legatură cu care se realizează predicţia este de tip calitativ (nominală sau ordinală) sau este cantitativă cu valori discrete;

 Algoritmi de regresie - când variabila în legătura cu care se realizează predicţia este cantitativă continuă (ia valori reale).

Clasificarea reprezintă procesul prin care se caută proprietăţi comune în seturi de obiecte din clase de date şi se clasifica în clase diferite în conformitate cu un model de clasificare. Clasificarea permite crearea modelelor pentru prezicerea membrilor unei clase. Scopul clasificării este în primul rând analiza datelor antrenate şi dezvoltarea pe baza acestor date a unui model, o descriere exactă pentru fiecare clasă folosind trăsăturile disponibile ale datelor. Clasificatorul

In the case of clickstream analysis is working on a database with server sessions that record user requests. User sessions are sequences S = (s1 ,..., sn) formats with the pages visited by the user. Determination of frequent links and association rules is essential for the clickstream analysis problem, how users navigate the Internet and accessing various sites.

Sequential Association Rules

Often, transactions are recorded taking into account a temporal sequence. For example, transactions for loyalty card holders correspond to sales receipts sequence. Transactions that record navigation paths followed by a web user is associated with a temporal sequence of sessions. In such situations, analysts are keen to extract association rules that take into account temporal dependencies. Sequential analysis is used to determine patterns of data using a temporal sequence of states. The problem of discovering sequential rules was first introduced by Agrawal and Srikant in [2].

Classification and regression are forms of supervised learning. Classification and regression are the largest category of applications, consisting of building models to forecast the membership to a set of class (classification) or to forecast of some values (regression). There are several techniques devoted to solving problems of classification and regression, including decision trees, Bayesian techniques, neural networks and k-NN enjoy wide recognition.

Supervised learning techniques aim to generate automatic induction mechanisms with predictive power by extracting information contained in the database and their transformation into a knowledge base.

There are two main classes of algorithms for induction:

 Classification algorithms - when the variable is done about that prediction is qualitative (nominal or ordinal) or quantitative with discrete values;

 Regression algorithms - when the variable about which the prediction is made is still quantitative continue (it takes real values). Classification is the process of seeking common properties from objects sets of class data and are classified into different classes according to a classification model. Classification allows you to create models to predict class members. The purpose of classification is primarily driven analysis based on these data and development of a model, an exact description of each class using the features of the available data. In order to be used the clasifier must first learn a mapping from a set of input variables and their values to predict output values for decision variables. Classifier can be used to predict output variables values using input

values once the pattern has been learned through the training data. Classification is often used in
pentru a putea fi folosit, trebuie mai întâi să învețe o cartografie dintr-un set de variabile de intrare şi valorile lor pentru a prevedea valorile de ieşire pentru variabilele de decizie. Clasificatorul poate fi folosit pentru a prezice valorile variabilelor de ieşire folosind valorile de intrare de îndată ce tiparul a fost învățat prin intermediul datelor de antrenament. În aplicații data mining în business clasificarea este intalnita des. De exemplu, clasificarea se întâlneşte în detectarea fraudei, unde clasificarea încearcă să identifice dacă tranzacția este legală sau suspectț. Alte exemple de utilizare a metodei de clasificare sunt definirea profilului cumpărătorului, analiza tratamentelor ineficiente, diagnosticarea medicala, aprobări de credite.

Clustering (Gruparea) este o formă de învățare nesupervizată care presupune căutarea bazelor de date de intrare de diferențe întâlnite între item-uri de date, descoperind astfel, în urma procesului de diferențiere, grupuri(clustere) de obiecte comune în datele de intrare. Clusterele sunt adesea folosite pentru schimbarea şi detectarea deviației în cadrul cărora scopul este găsirea item-urilor de date care nu se încadrează în normă, sau grup(cluster). Obiectele din acelaşi cluster trebuie să aibă profile similare (omogenitate intra-cluster), iar obiectele din clustere diferite să aibă profile distincte (eterogenitate inter-clustere). Schimbarea şi detectarea deviației se aplică într-o multitudine de domenii, precum este detectarea tranzacțiilor frauduloase (frauda de telefoane sau a cardurilor bancare), detectarea tratamentelor medicamentoase nepotrivite înainte de a fi prea tarziu, precum şi detectarea noilor tendinte de market. În e-business clustering este folositoare deoarece poate lucra cu colecții mari de date şi foloseste la realizarea diferitelor grupe pe baza caractericilor comune ale obiectelor. Poate fi folosită şi inaintea aplicării metodei de clasificare. De exemplu, dacă folosim metoda clustering pentru o listă de profile ale utilizatorilor, un cadru (schelet) al diferitelor tipuri de clienți poate fi construit. Această metodă de clustering are aplicații diverse în: marketing, suport clienți şi determinarea fraudelor (daca comportamentul unui utilizator de telefon celular sare imediat de la un cluster la altul, aceasta poate indica un jaf de telefon sau o clonare).

Etapele procesului de clustering

Etapele unui process de clustering presupun rezolvarea următoarelor probleme:

 Stabilirea elementelor supuse procesului de clustering este o etapă principală care uneori mai include şi stabilirea numărului de clase/grupe, tipul şi scara caracteristicilor/atributelor disponibile algoritmului de clustering.

 Selectarea caracteristicilor reprezintă procesul de identificare a celor mai utile attribute /caracteristici utilizate în procesul de clustering. Se referă la o modalitate de a efectua una sau mai

values for decision variables. Classifier can be used to predict output variables values using input values once the pattern has been learned through the training data. Classification is often used in business data mining applications. For example, the classification meets in detecting fraud, where classification is trying to identify if the transaction is legal or suspect. Other examples of using the method of classification is to define customer profile analysis of ineffective treatments, medical diagnosis, credit approvals.

Clustering is a form of unsupervised learning which involves searching databases for input differences found between the items, and found, in the process of differentiation, groups (clusters) of objects in the input data. Clusters are often used to change and detect of deviation aimed at finding items have data that does not fit the norm, or group (cluster). Objects in the same cluster should have similar profiles (intra-cluster homogeneity) and objects in different clusters have distinct profiles (inter-cluster heterogeneity). Change and deviation detection is applied in many fields, such as is detecting fraudulent transactions (fraud phones or bank cards), detect inappropriate drug treatment before it is too late and detect new market trends. In e-business clustering is useful because it can work with large collections of data and uses the achievement of different groups based on common objects features. Can be used before applying the method of Classification. For example, if we use the clustering method for a list of user profiles, a framework of different types of clients can be built. This clustering method has various applications in marketing, customer support and determination of fraud (if the behavior of a cell phone user immediately jumps from one cluster to another, this may indicate a phone robbery or cloning).

Clustering process involves stages of solving the following problems:

 Lay the subject of clustering process is a main stage which sometimes includes setting the number of classes / groups, type and scale characteristics / attributes available clustering algorithm.

 Feature extraction is the process of identifying the most useful attributes/ features used in the clustering. It refers to a way to make one or more transformations of input data in order to obtain new dominant features .

multe transformări ale datelor de intrare în scopul obţinerii unor noi caracteristici dominante.

 Definirea unei măsuri de proximitate în cadrul unei mulţimi. Proximitatea elementelor este măsurată printr-o funcţie de distanţă definită pe perechi de elemente. Măsurile de asemanare pot fi folosite şi pentru a caracteriza similitudinea conceptuală dintre doua sau mai multe elemente.

 Procesul de clustering poate fi realizat în mai multe feluri. Datele de ieşire pot fi „hard" (separarea elementelor în grupe clar determinate) sau fuzzy (în care fiecare element are un grad variabil de apartenenţă la fiecare din grupele rezultate)

 Extragerea rezultatelor reprezintă procesul de obţinere a rezultatelor într-o formă cât mai simplă şi reprezentativă. Extragerea rezultatelor reprezintă o descriere concisă a fiecarei grupe obţinute, de obicei prezentate sub formă unor elemente reprezentative. Toţi algoritmii de clusterizare ar trebui să conducă la obţinerea unor grupe/clase pentru orice mulţime de date de intrare. Dacă în urma procesului de clustering folosind un anumit algoritm nu se obţine gruparea elementelor, atunci se aplicţ un alt algoritm care poate furniza rezultate mai bune decat cel anterior.

 Analiza validităţii grupelor efectuează o evaluare a rezultatelor procesului de clustering, de obicei un criteriu de optimizare. Se verifică dacă rezultatele grupării spaţiale sunt corecte.

4. Aplicaţii ale metodelor data-mining în e-business

Marketingul Direct. Datorită dimensiunii şi complexităţii pieţei actuale, marketing-ul de masă a devenit tot mai scump, neprofitabil fiind înlocuit de marketingul direct, care se bazează pe selectarea grupurilor ţintă de clienţi şi stabilirea de interacţiuni individuale cu aceştia pe multiple canale de comunicare. Astfel, companiile se repoziţioneaza strategic, orientarea produs-centrică se transformă rapid în una client centrică.

Managementul relaţilor cu clienţi (CRM) are ca obiect elaborarea de strategii pentru atragerea de noi clienţi, menţinerea celor existenţi şi recâştigarea celor care au migrat către alţi ofertanţi. La nivel operaţional, CRM cuprinde toate activităţile ce privesc contactul direct cu consumatorul. La nivel analitic, CRM furnizează o serie de metode pentru analiza comportamentului clienţilor prin analiza datelor obţinute prin sistemele de procesare a tranzacţiilor.

Managementul analitic al relaţiilor cu clienţii are trei obiective majore :

 Segmentarea pieţei , care reprezintă procesul de împărţire a clienţilor în grupe cât mai omogene intern pe baza similarităţilor manifestate (obiceiuri, gusturi, afinităţi), aceste grupe fiind cât mai eterogene între ele. Astfel, firma poate trata personalizat diverse segmente de clienţi şi se poate

 Defining a measure of proximity in a crowd. The proximity of elements is measured by the distance function defined on pairs of elements. Similarity measures can be used to characterize the conceptual similarity between two or more items.

 Clustering process can be accomplished in several ways. Output data can be hard (separation of elements in clearly defined groups) or fuzzy (in which each element has a variable degree of membership of each group results)

 Extraction of results is the process of obtaining results in a simpler form and representative. Extraction results is a concise description of each group obtained, usually in the form of representative elements. All clustering algorithms should lead to the achievement of groups / classes for any set of inputs. If in the process of using a clustering algorithm does not get group items, then apply another algorithm that can provide better results than the previous.

 Validity analysis group performed an evaluation of clustering process, usually a criterion for optimization. It is checked if the results of spatial clustering are correct.

4. Applications of data mining methods in e-business

Direct Marketing. Due to the size and complexity of the current market, mass marketing has become increasingly expensive, unprofitable, so being replaced by direct marketing, which is based on selecting target groups of clients and establishing individual corelations with them on multiple channels. Thus, companies strategic are repositioned, product-centric orientation quickly transforms to a client centric.

Customer Relationship Management (CRM) target is to develop strategies to attract new customers, maintain existing ones and regaining those who migrated to other bidders. From operational point, CRM includes all activities relating to direct contact with the consumer. At the analytical level, CRM provides a number of methods for analyzing customer behavior by analyzing data obtained through transaction processing systems.

Analytical customer relationship management has three major objectives:

 Market segmentation, which is the division of customers into homogeneous groups based on the internal as manifested similarities (habits, tastes, affinities), this group is more heterogeneous among themselves. Thus, the firm may treat different segments of customers personalized and can be concentrated on certain target groups that correspond to some criteria of profitability.

 Consumer profiling involves modeling consumer behavior based on a wide range of attributes such as the geographical, cultural and ethnic, economic conditions, frequency of

concentra prioritar asupra anumitor grupuri ţintă ce corespund anumitelor criterii de profitabilitate.

 Stabilirea profilului consumatorului presupune modelarea comportamentului consumatorilor în funcţie de o paletă largă de atribute precum sunt cele geografice, culturale şi etnice, condiţii economice; frecvenţa de cumpărare, frecvenţa plângerilor şi reclamaţiilor, preferinţele şi gradul lor de satisfacere; vârstă, educaţie, stilul de viaţă, canalele media utilizate, metoda de recrutare la care a raspuns clientul.

 Poziţionarea produsului în preferinţele potenţialilor clienţi este un instrument de marketing centrat pe identificarea celor mai atractive trăsături ale unui produs astfel încât să maximizeze tentaţia cumpărătorului de a-l cumpăra. Aici apare aşa numita problemă a coşului de cumpărături. Se determină probalilitatea ca anumite produse să fie cumpărate împreună.

5. Concluzii

În lumea afacerilor de azi, folosirea calculatorului pentru procesul de business şi înregistrarea datelor a devenit omniprezent. Odată cu apariţia acestei vârste electronice vine şi un produs nepreţuit-datele (informaţiile). Virtual, fiecare mare companie işi înregistrează toate tranzacţiile.

Data mining este procesul folosit pentru a lua acest imens volum de date şi a le transforma în cunoştinţe folositoare. Data Mining se referă la procesele de selectare a unor relaţii necunoscute anterior cu scopul obţinerii unui rezultat curat şi folositor celui care deţine baza de date.

Drept rezultat, un sistem data minig are câteva faze. Fazele prezentate încep cu randul de date şi se încheie cu extragerea cunoştinţelor care s-a produs ca urmare a parcurgerii etapelor: selectarea, preprocesarea, transformarea, data mining, interpretarea şi evaluarea. Originile tehnicilor data-minig au fost gândite ca venind din trei arii ale învăţării şi cercetarii : statistică, învăţarea maşinilor şi inteligenţă artificială. Prima fundaţie a metodelor data mining a fost in statistica. Statistica este baza majorităţii tehnologiilor pe care se bazează data minig. Multe din domeniile statisticii, precum analiza regresiei, distribuţii standard, deviaţii şi variaţii standart, analiza grupului sunt construcţiile tehnice mai avansate ale statisticii care stau la baza tehnicilor data-minig.

Pentru a se diferenţia în cadrul economiei pe internet, întreprinderile învingătoare trebuie sa realizeze că e-business înseamnă mai mult decât simple tranzacţii de cumpărare/vânzare, strategiile corespunzătoare fiind cheia succesului pentru a îmbunătăţi puterea de competiţie. Acest lucru se

purchase, frequency of complaints, preferences and their degree of satisfaction, age, education, lifestyle, media used, method of recruitment that the customer response.

 Positioning the product in the preferences of potential customers is a marketing tool focused on identifying the most attractive features of a product to maximize customer temptation of buying it. Hence the so-called problem of shopping basket analysis. It determines the probability that certain products are purchased together.

5. Conclusions

In today's business world, computer use for business process and data recording has become ubiquitous. With this electronic age comes an invaluable product-data (information). Virtually every large company records all its transactions.

Data mining is the process used to make this huge volume of data and turning them into useful knowledge. Data Mining refers to the process of selection of previously unknown relationships in order to obtain a clean and useful result to that which holds the database.

As a result, a data minig system has several phases. Phases begin to turn data set and ends with knowledge extraction that occurred as a result of carrying out the steps: selection, preprocessing, transformation, data mining, interpretation and evaluation.

The origins of data minig techniques were designed as coming from three areas of learning and research: statistical, machine learning and artificial intelligence. The first foundation of data mining methods was in statistics. Statistics is the most technology that relies on data minig. Many of the statistics domains such as regression analysis, standard distributions, standard deviations and variations, the group analysis are construction techniques of advanced statistical techniques underlying data minig.

To differentiate into the Internet economy, companies must realize that winning e-business means more than simple transactions of purchase / sale, appropriate strategies are the key to improve competitive power. This can be done using data mining techniques and other statistical analysis on historical data from e-business activities.

poate realiza utilizând tehnici data mining, precum şi alte analize statistice pe datele istorice rezultate din activităţile e-business.

Referinţe:

[1] Adam Jolly, (2003), *The Secure Online Business,* Kogan Page and Contributors.
[2] Agrawal, R., Srikant, R. (1995), Mining sequential patterns, *International Conference on Data Engineering(ICDE'95)*, Taipei, Taiwan, pp. 3-14.
[3] Award, Elias, *Electronic Commerce from Vision to Fulfillment,* Pearson Education, Upper Saddle River, New Jersey, 2002.
[4] Berry, M., Linoff, G. (1997), Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons, Chichester.
[5] Claudia Elena Dinucă, *E-Business, a new way of trading in virtual environment based on information technology*, Annals of the "Ovidius" University, Economic Sciences Series Volume XI, Issue 1 /2011
[6] Dunham, M.H. (2003), Data Mining : Introductory and Advanced Topics. Prentice Hall, Pearson Education Inc.
[7] Gunjam Santami, (2002) *B2B Integration –A Practical Guide to Collaborative E-commerce,* Imperial College Press: London.
[8] Harmon, P; Rosen, M; Guttman, M (2001) *Developing E-Business Systems & Architectures- A Manager's Guide*; SUA: Academic Press.
[9] Janice Reynolds (2004) , *The Complete E-Commerce Book: Design, Build, & Maintain a Successful Web-based Business*, Second Edition, CMP Books.
[10] Jatinder N.D. Gupta and Sushil K. Sharma Ball şi alţii *Intelligent Enterprises of the 21st Century* (2004) SUA: Idea Group.
[11] Jiawei Han, Micheline Kamber (2006), *Data Mining Concepts and Techniques* Second Edition*,* USA: Elsevier.
[12] Mike Havey **,** *Essential Business Process Modeling* (2005), SUA: O'Reilly.
[13] Nong, Y. (2003), The handbook of Data Mining, Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey.
[14] Porter, Michael E. *Competitive Strategy: Techniques for Analysing Industries and Competitors.*
[15] Raisinghani, M (2004) *Business Intelligence in the Digital Economy: Opportunities, Limitations, and Risks* ; SUA:Idea Group Publishing.
[16] Turban, Efraim; King, David, *Introduction to E-commerce*, Pearson Education, Upper Saddle River, New Jersey, 2003.
[17] Vercellis, C. (2009), *Business Intelligence: Data Mining and Optimization for Decision Making*,UK: John Wiley & Sons.

References:

[1] Adam Jolly, (2003), *The Secure Online Business,* Kogan Page and Contributors.
[2] Agrawal, R., Srikant, R. (1995), Mining sequential patterns, *International Conference on Data Engineering(ICDE'95)*, Taipei, Taiwan, pp. 3-14.
[3] Award, Elias, *Electronic Commerce from Vision to Fulfillment,* Pearson Education, Upper Saddle River, New Jersey, 2002.
[4] Berry, M., Linoff, G. (1997), Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons, Chichester.
[5] Claudia Elena Dinucă, *E-Business, a new way of trading in virtual environment based on information technology*, Annals of the "Ovidius" University, Economic Sciences Series Volume XI, Issue 1 /2011
[6] Dunham, M.H. (2003), Data Mining : Introductory and Advanced Topics. Prentice Hall, Pearson Education Inc.
[7] Gunjam Santami, (2002) *B2B Integration –A Practical Guide to Collaborative E-commerce,* Imperial College Press: London.
[8] Harmon, P; Rosen, M; Guttman, M (2001) *Developing E-Business Systems & Architectures- A Manager's Guide*; SUA: Academic Press.
[9] Janice Reynolds (2004) , *The Complete E-Commerce Book: Design, Build, & Maintain a Successful Web-based Business*, Second Edition, CMP Books.
[10] Jatinder N.D. Gupta and Sushil K. Sharma Ball and other *Intelligent Enterprises of the 21st Century* (2004) SUA: Idea Group.
[11] Jiawei Han, Micheline Kamber (2006), *Data Mining Concepts and Techniques* Second Edition*,* USA: Elsevier.
[12] Mike Havey **,** *Essential Business Process Modeling* (2005), SUA: O'Reilly.
[13] Nong, Y. (2003), The handbook of Data Mining, Lawrence Erlbaum Associates, Publishers Mahwah, New Jersey.
[14] Porter, Michael E. *Competitive Strategy: Techniques for Analysing Industries and Competitors.*
[15] Raisinghani, M (2004) *Business Intelligence in the Digital Economy: Opportunities, Limitations, and Risks* ; SUA:Idea Group Publishing.
[16] Turban, Efraim; King, David, *Introduction to E-commerce*, Pearson Education, Upper Saddle River, New Jersey, 2003.
[17] Vercellis, C. (2009), *Business Intelligence: Data Mining and Optimization for Decision Making*,UK: John Wiley & Sons.