Въведение

Терминът Data Mining в превод означава “добиване” или “извличане на данни”. С него се свързват и термините ”откриване на знания в базата от данни ”(Knowledge discovery in databases) и ”интелектуален анализ на данните”. Възникването на всички тези термини е свързано с появата на нови насоки в разработването на средства и методи за обработка на данните. Data mining представлява процес за анализ на съхраняваните бази данни в посока на извличане на нова полезна информация чрез разкриване на дълбоките и скрити взаимоотношения между на пръв поглед неизвестни и несвързани една с друга величини. Важна негова особеност е, че той осигурява възможност за обработка на многомерни масиви и извличане на многомерни зависимости като същевременно автоматично разкрива изключителните ситуации - данни и случай не включващи се в общите закономерности. Data mining анализа автоматически прави хипотези за разкриване на зависимости между различни компоненти и параметри. Работата на аналитиците, които се занимават с тези системи се свежда до проверка и доуточняване на получените хипотези. Появата на Data mining е свързана с необходимостта от усъвършенстване на техниките за запис и съхранение на данните, които обобщават труда на хиляди хора в колосални потоци от информация в най-различни области. С времето е станало ясно, че без продуктивна обработка на данните се правят никому ненужни извадки. Нуждата в развитието на съвременните технологии от такава преработка на данните може да се обобщи в следното:

- Неограниченият обем на данните.
- Голямата разнообразие и разнородност на данните (количествени, качествени и текстови).
- Необходимост от конкретни и разбираеми резултати.
- Инструменти за обработка на данните предоставящи възможност за лесно използване.

В основата на съвременните технологии Data mining стой концепцията за шаблони или модели, отразяващи фрагментираните многоаспектни взаимоотношения между данните. Тези шаблони представят сбор от закономерности, подбор на данните по дадени свойства, които са подходящо представени във форми лесно достъпни за потребителите. За създаването на тези шаблони се прилагат методи, които не ограничават основното предположение в структурата на модела и вида на разпределените значения на анализираният показател.

Характерна особеност на Data mining анализа е, че той е тясно свързан с OLAP системите, но между двата метода има и принципни различия, които могат да се видят от таблицата:

Таблица 1. Примерно формулирани задачи при методите на OLAP и Data Mining

OLAP	Data Mining
Какъв е средният показател на заболяванията на пушачите и непушачите?	
Среща ли се точен шаблон при описаните случаи потвърждаващи повишената заболеваемост	
Какъв е средния размер на телефонните услуги ползвани от настоящите абонати в сравнение	
Какво охарактеризира абонатите, които биха се отказали от услугите на телефонната компани	
Каква е средната величина на покупките направени с откраднатите или неоткраднатите креди	
Каква е схемата на покупките направени с откраднати кредитни карти?	

Важно предимство на Data mining анализа е непредвидимостта в издирените шаблони. Това означава, че откритите шаблони трябва да отразяват неочевидни, неочаквани зависимости в данните представляващи част от т. нар. скрити значения. Поради това е дошла идеята, че “необработените” данни съдържат много по-дълбоки пластове от скрити знания, които могат да бъдат разкрити само при едно детайлно проучване в дълбочина. Това е представено в таблица 1.

От таблицата се вижда, че Data mining извлича дълбоко скрити данни, които чрез OLAP не могат да бъдат разкрити и анализирани, като търсенето става отгоре надолу. Понятието OLAP (Online Analytical Processing) обхваща технологията за многомерен анализ, която позволява използуването на информацията съхранена в data warehouse. Обикновено тя включва средства за интерактивен анализ на данните, които се извличат от различни бази и се обобщават за нуждите на даден потребител. OLAP средствата предоставят възможности за представяне на данните в различни разрези, поради което са значително по-сложни от традиционните релационни бази от данни. От своя страна Data mining също се използува за анализ на данните, но обхваща технологии, позволяващи да се открият неявни шаблони и взаимодействия в различни извадки. Съществуват и т.нар. Data Marts, които съхраняват подмножества от агрегирани данни и могат да се разглеждат като локални Data Warehouses. Информацията получена от Data mining може да се използва след това за увеличаване на фирмената ефективност. Например при анализирането на потребителските особености на потребителите, дава възможност да се предскаже поведението на потребителите и да се повлияе върху него.

Сфера на приложение на Data mining анализа.

Кръгът му на приложение е много голям. Той се използва навсякъде, където има нужда от обработване на данни. А данни и то в големи количества се срещат навсякъде в съвременната дейност на човека. Но още при първата си поява Data mining и неговите методи са заинтригували предимно комерсиалните търговски предприятия, разработващи проекти на основата на поддържането на информационни хранилища от данни. Опитът на много такива предприятия показва, че ползата от Data mining анализа може да достигне до – 100 %. Data mining е и много ценна за много от ръководителите и аналитиците в изпълняването на всекидневната им дейност. Чрез Data mining анализа те могат да получат още едно предимство в силната пазарна конкуренция на бързо развиващото се общество. Приложение на Data mining в различните отрасли могат да се разделят на:

Бизнес приложения:

- Разносна търговия - проявление на товарите, които се разполагат съвместно, избор на местоположение на товарите в магазина, анализ на потребителските нужди, прогнозиране на избора.
- Маркетинг - изграждане на различни сегменти с потребители, тенденции в покупателното поведение.
- Финанси - проявление на правила за експертни системи за андеррайтинг (under writing), класификация на дебиторските задължения по възможни вземания, прогнози в измененията на валутния курс.
- Здравеопазване - анализ на резултатите от лечението на пациентите, анализ на контактите.
- Промишлено производство - диагностициране на неизправностите.

Неикономически приложения:

- Медицина - използва се за поставянето на медицински диагнози, чрез съчетаването на различните симптоми.
- Молекулярна генетика - разкрива закономерности в експерименталните данни. Те дават възможност за предсказване на последствията от промените в генетичния код на живите организми и неговото разкодиране.
- Органична и неорганична химия - разглеждат се особеностите на химичният строеж на веществата, съединенията, в които участват и свойствата им.

Data mining анализа може да се разгледа в различни аспекти, но тук ще се спрем на примери от неговото бизнес и в частност маркетингово приложение.

Тя може да помогне на дадена фирмата точно да планира и оцени своята дейност. За пример можем да вземем анализа на потребителското поведение. Той дава възможност да се анализират предпочитанията на потребителите така, че най-добре да се удовлетворят те в съответствие с доходите им. Потребителското поведение присъства в събраните данни в скрит вид и за да се разкрие е необходимо да се използва Data mining. След това може да се определи, че клиентите, които искат да закупят стока А заедно с нея биха закупили и стока Б. Тази информация може да се реализира чрез маркетинговата стратегия на фирмата: тази стока може да се постави на витрината, точно до другата, която клиента ще закупи със сигурност. Така по-бързо ще се увеличи оборота на фирмата и ще се повишат и печалбите. Маркетинговото приложение на Data mining може да се представи чрез:

- Моделиране на потребителското поведение - основано е на демографски показатели и история на продажбите. Помага да се определи как клиентите биха реагирали на даден продукт или рекламна кампания.
- Оценка на потребителите - основава се на често повтарящите се покупки, похарчени суми и продължителност на сътрудничеството. Позволява да се изясни, кои са най-ценните за фирмата клиенти.
- Сегментация на потребителите - какви общи характеристики има между основните клиенти на фирмата и възможно ли е те да се групират в определени групи.
- Планиране на продажбите - ако има информация, че клиентите закупуват продукт А, Б или В, то каква е вероятността да закупят продукт Г.

Можем да обобщим, че Data mining анализа и намира приложение в области, където не са достатъчни само статистически и аналитични методи за изграждане на подходящите модели. В тези области преобладават нееднородни, хетерогенни, променливи и в големи количества данни.

Технологии и алгоритми, използвани в Data mining анализа.

В основата на Data mining анализа стоят две технологии: машинно обучение и визуализация - визуално представяне на информацията. Качеството на визуализацията определя графическото представяне на данните в техните цветове, форми и други елементи представящи скритите връзки между данните. Ефективността на методите за машинно обучение се определя от възможностите на Data mining анализа да изследва много по-големи количества данни и да разкрива връзките между тях, отколкото може човек. Машинното обучение предполага използване на различни методи:

- Дърво на решенията – явява се един от най-популярните подходи, тъй като се характеризира с нагледност и разбираемост. Но дървото на решенията принципно не може да намери най-добрите пълни и точни правила в данните. Предназначен е за класифициране на данните като използват тежестта на коефициентите на разпределение на елементите на данните във все по-малки и по-малки групи.

Голяма част от системите използват именно този метод. Най-известни са ee5/C5.0 (RuleQuest, Австралия), Clementine (Integral Solutions, Великобритания), SIPINA (University of Lyon, Франция), IDIS (Information Discovery, США), KnowledgeSeeker (ANGOSS, Канада).

- Асоциативни правила - този метод класифицира данните на основата на набор от правила, които са подобни на експертните системи. Като тези правила могат да се генерират чрез използването на процес по изискване и проверка на различни комбинации от правила или на извличането на такива от дървото на решенията.
- Генетически алгоритми - чрез него се определят естествени “разбивки “ на данните основани на целеви променливи. Всеки клон на дървото е отделна част от правилата.

- Нейронни мрежи - при този метод знанията се представят във вид на връзки, съединяващи набор от условия. Силата на връзката се определя от отношението между факторите и данните.

В таблица 2 е представено кратко описание на основните алгоритми в data maining.

Таблица 2

алгоритми	описание
Асоциативни правила	

Представят причинно следствената връзка и определят вероятността или коефициента на до

Дърво на решенията и алгоритъм на класификацията

Определя естествени “разбивки “ на данните основани на целеви променливи. Първоначално с

Изкуствени нейронни мрежи

Тук за предсказване на значението на целевия показател се използват набор от входни проме

Генетически алгоритми

Този метод използва интерактивният процес в еволюцията на последователните поколения мо

Извод на базата на съпоставяне (Memory-based Reasoning, MBR) или извод на базата на преце

Тези алгоритми са основани на определяне на аналогии, които са най-близко до текущата ситу

Клъстерен анализ.

Подразделяне на хетерогенните данни на хомогенни или полухомогенни групи. Този метод поз

Всеки метод има своите преимущества и недостатъци. Преимуществото на дървото на решенията и асоциативните правила се състой в тяхната разбираемост - те приличат на естественият език. Недостатъкът им е че не са подходящи за много широки числови интервали. Това се дължи на факта, че всяко правило в дървото на решенията представлява една връзка, зависимост или отношение. Преимуществото на нейронните мрежи се състой в компактното представяне на числовите отношения за широк диапазон от значения. Недостатък - сложностите в интерпретирането му.

Типове закономерности на Data Mining.

Има пет типа закономерности, които позволяват да се реализира Data Mining анализа:

- *Асоциация.* Прилага се в случаите, когато няколко събития са свързани едни с други, например изследване проведено в супермаркетите може до покаже, че 65% от купуващите пуканки си купуват и Кока кола, а при наличието на отстъпка за такъв комплект покупките на Кока кола се увеличават с 85%. Разполагайки със сведения за подобна асоциация, мениджърите лесно мога да преценят колко процента да бъде тази отстъпка.

- *Последователност*. Ако съществува верижност по време на събитията се говори за последователност. Така например след покупката на жилище в 45% от случаите в течените на месец се закупува и кухненско обзавеждане, а след това и 60% закупуват и хладилник.

- *Клъстеризация*. С нейна помощ от класификационните множества се извличат хомогенни (еднородни) групи от данни имащи сродни признаци. Тя разширява възможностите за прогнозиране.

- *Класификация*. С нейна помощ се разкриват признаци, характеризиращи групите, в които се включва даден обект. Това става посредством анализ на класифицируемите обекти и формулиране на определен набор от правила.

- *Прогнозиране*. В основата на съвременните прогнози в технологията на Data mining анализа стоят данните намиращ се в хранилищата от данни (Data warehouse). Въз основа на тях се построяват шаблони, отразяващи динамиката на поведението на целевият показател, с чиято помощ може да се предскаже поведението на системите в бъдеще. Data warehouse (хранилищата от данни) се дефинира като множество от интегрирани, тематично ориентирани бази от данни, проектирани за поддържане на процеса "вземане на решения", където всяка единица от данни е смислена в определен момент от време . Този информационен масив съдържа както самостоятелни и така силно обобщени данни. На фигура 3 е представено мястото на Data Mining анализа сред традиционните компоненти на Data Warehouse, както и взаимодействията между тях.

Инструменти на Data Mining анализа и разработването на Data Mining приложения.

Съществува широк кръг от инструменти за поддържането на Data Mining анализа. Тук се отнасят, както общо достъпните алгоритми за визуализация и машинно обучение, така и сложни програмни пакети, работещи на базата на паралелните процесори, като тяхната стойност може да достигне до няколко хиляди долара. Използването на най-подходящият инструмент за извършването на Data Mining анализа се определя от ред на условията и целите на проекта например анализа на потребителската кошница. При избора на инструменти или алгоритми е много важна гъвкавостта - доколко чрез избраната стратегия може да бъде получен желаният резултат. Разработването на Data Mining приложения в сферата на бизнеса преминава през няколко стъпки:

Стъпка 1: Установяват се мащабите на проекта, определящи какви данни е необходимо да се съберат. Важно е проектът да бъде направен, така че да решава конкретни бизнес цели.

Стъпка 2: Разработване на бази данни за Data Mining. Необходимата информация може да бъде разпределена по няколко бази данни, които понякога могат и да не са в електронна форма. Данните между различните приложения е необходимо да се консолидират и обобщят, за да се премахнат несъответствията. Развитието на Data Mining - анализа не трябва да променя алгоритмите, които са свързани с изработването на витрини от данни - извадки от базата данни по определено свойство. Фактически за ефективен анализ трябва да има наличие на корпоративно хранилище от данни, което излиза много по-евтино отколкото използването на отделни витрини от данни. С внедряването на Data Mining проекти в предприятието количествените ползи растат, но и възниква необходимостта от осигуряване на достъп до корпоративните структури от

данни. Съвременните хранилища от данни представляват не само ефективен способ за съхранение на всички корпоративни данни, но и представляват идеална основа за разработването на Data Mining проекти. Складирането на данните в предприятието обезпечава съгласуването и актуализирането на данните с тези на клиентите. Внедряването на Data Mining функциите в хранилищата от данни съкращава с два пъти разходите. В този случай, първо не е необходимо да се закупува и обслужва допълнително оборудване за Data Mining. Второ за компанията не е необходима да пренася данните от хранилищата до специални източници, които после ще се използват от Data Mining - така се икономисва време и материални ресурси. Още един важен момент е изтриването на данните. Тук се включва проверка на целостта и обработка на съществуващите значения. Точността на метода Data Mining зависи от качеството на информацията избрана за негова основа.

Стъпка 3: Даване на количествена оценка на елементите от данни. Като например: Кога един човек може да се нарече “разточителен”- тогава когато харчи 50 или 300 лв. на седмица? Сътрудничеството с експериментите в предметната област помагат да се решат такива или подобни въпроси и да се отделят елементите от данни, които осигуряват най-голям смисъл за нуждите на бизнеса.

Стъпка 4: Примерни алгоритми на Data Mining за определяне на отношението между данните. Не е изключено, за получаването на нужните зависимости да се използват няколко различни алгоритми. Едни могат да се използват в началото на процеса а други в краят му. Понякога могат да се използват и няколко паралелни алгоритми, за да се получат данни с различна точност.

Стъпка 5: Изследване на съотношенията проявили са на предходния етап от прилагането на проекта. На този етап може да потрябва помощта на експерт в съответната предметна област. Той определя дали тези съотношения са специфични или общи и указва в каква област трябва да продължи анализа.

Стъпка 6: Представяне на резултатите във вид на отчет, в който да са разкрият преизчисленията за всички интерпретирани отношения. Такъв отчет донася изгода тогава, когато експертът може да приложи творчески подход при анализирането на данните и ползите от тях. Като след това фирмата разработчик е длъжна не само да научи клиента на методиката на поисканата от него зависимост в данните, но и да се обърне особено внимание на обучението за работа с програмата. Целият първи прототип на проекта се състои в това да се намали количеството на грешките в базата от данни (в първи, втори, трети и пети етап). За да се достигне до всички тънкости в

изследваните данни трябва да се направят няколко итерации или заместявания и тогава да се предаде окончателният проект. При разработването на Data Mining проекта влияят и други фактори: типът на крайното приложение; наличие и състояние на хранилищата от данни; сроковете, в които трябва да се завърши проекта; обемът на данните, тяхното разнообразие и характеристики.

Връзка на Data Mining с други области.

Data Mining анализа е мултидисциплинарна област. Тя включва съвкупност от знания и разработки в различни области на човешката дейност като:

- Информационен анализ
- Оперативно аналитична обработка на данните - OLAP
- Бази данни
- Data Warehouses - хранилища за данни.
- Ефективни изчисления
- Статистика
- Визуализиране на данните
- Разпознаване на образите
- Нейронни връзки
- Експертни системи

Система за визуализация на данните получени от Data Mining анализа.

Тя заема важно място във всеки Data Mining проект. Тя осигурява графическо представяне на получените данни, като графики, диаграми, схеми, таблици и др.Това става като системата за визуализация поддържа дружелюбен интерфейс, позволяващ лесно асоцииране на анализираните показатели с различните параметри на диаграмите като цвят, фон, форма, ориентация спрямо основните оси, размер и др. Системата за визуализация трябва да предоставя и собствени средства за мащабиране за по детайлно разглеждане. Характерно за тези системи е че са доста скъпи.

Съвременните тенденции в Data Mining анализа - Deep Data Mining (DDM).

Съвременните тенденции на пазара сочат, че приложението и използването на Data Mining технологията непрекъснато нараства и се развива. Фирмите разработчици се ориентират бързо към запълването на тази ниша на пазара чрез предлагането на нови приложения. Като ново предложение можем да разгледаме Deep Data Mining (DDM). Това е нова технология, основаваща се на специална локална геометрия. В тази геометрия всеки обект съществува в собствено локално пространство и събитие със собствени размери. При всяко поискване на логически закономерности в данните между локалните процедури се получава геометрично тълкувание. Технологията DDM дава възможност да се разкрият данните чрез IF THEN правила, включващи десетки хиляди съвместно засичащи се логически събития, характерни за една съвкупност от данни и не характерни за останалите. Получават се резултати с много по-голяма ефективност и много по-близки до желаните. Съществено допълнение към новата технология се явява използваната нова формула ”данни + шум”. Тя е нововъведение в анализа на специални фалшификатори - обекти, осигуряващи равномерност във вероятностният смисъл на разпределените в пространството събития. Фалшификаторите представляват резултати, които са близки до желаните, но не ни интересуват и е добре да се игнорират. Делът на тези фалшификатори може да съответства или дори да превишава дела на изходната информация. Стълкновението на данните с фалшификаторите спомага за по-голяма устойчивост на получените логически закономерности и резултати.

Заключение

Като заключение можем да кажем, че Data Mining анализа намира приложение в тези области, където не са достатъчни само статистическите и аналитичните методи и изгражданите чрез тях модели. Data Mining анализа е подходящ за области, където преобладават нееднородни, хетерогенни, нестационарни, променливи и големи по обем данни.Това става при всички компании, които извършват обработка на данни при осъществяването на ежедневните си дейности и операции.