ГОЛЕМИ ДАННИ И ИЗВЛИЧАНЕ НА ЗНАНИЯ ОТ ДАННИ – ОСНОВНИ КОНЦЕПЦИИ

Информационната ера

Живеем в свят, в който ежедневно се събират огромни количества данни. Анализирането на такива данни е важна необходимост. „Живеем в ерата на информацията“ е популярна поговорка; но всъщност живеем в ерата на данните. Терабайти или петабайти1 данни се изливат в нашите компютърни мрежи, световната мрежа (WWW) и различни устройства за съхранение на данни всеки ден от бизнеса, обществото, науката и инженерството, медицината и почти всеки друг аспект от ежедневието. Този експлозивен ръст на наличния обем данни е резултат от компютъризацията на нашето общество и бързото развитие на мощни инструменти за събиране и съхранение на данни.

Бизнесите по целия свят генерират гигантски набори от данни, включително транзакции за продажби, записи за борсова търговия, описания на продукти, промоции за продажби, фирмени профили и представяне, както и обратна връзка с клиентите. Например големи магазини, като Wal-Mart, обработват стотици милиони транзакции на седмица в хиляди клонове по целия свят. Научните и инженерните практики генерират високи порядъци от петабайти данни по непрекъснат начин, от дистанционно наблюдение, измерване на процеси, научни експерименти, производителност на системата, инженерни наблюдения и наблюдение на околната среда.

Глобалните опорни телекомуникационни мрежи пренасят десетки петабайта трафик на данни всеки ден. Медицинската и здравната индустрия генерира огромни количества данни от медицински досиета, наблюдение на пациенти и медицински изображения. Милиарди търсения в мрежата, поддържани от търсачки, обработват десетки петабайти данни дневно. Общностите и социалните медии стават все по-важни източници на данни, произвеждащи цифрови снимки и видеоклипове, блогове, уеб общности и различни видове социални мрежи. Списъкът с източници, които генерират огромни количества данни, е безкраен.

Този експлозивно нарастващ, широко достъпен и гигантски масив от данни прави нашето време наистина ерата на данните. Мощни и гъвкави инструменти са силно необходими за автоматично разкриване на ценна информация от огромните количества данни и за трансформиране на тези данни в организирано знание. Тази необходимост доведе до раждането на извличането на данни. Сферата е млада, динамична и перспективна. Извличането на данни има и ще продължи да прави големи крачки в нашето пътуване от ерата на данните към идващата информационна епоха.

Пример 1.1 Извличането на данни превръща голяма колекция от данни в знания.

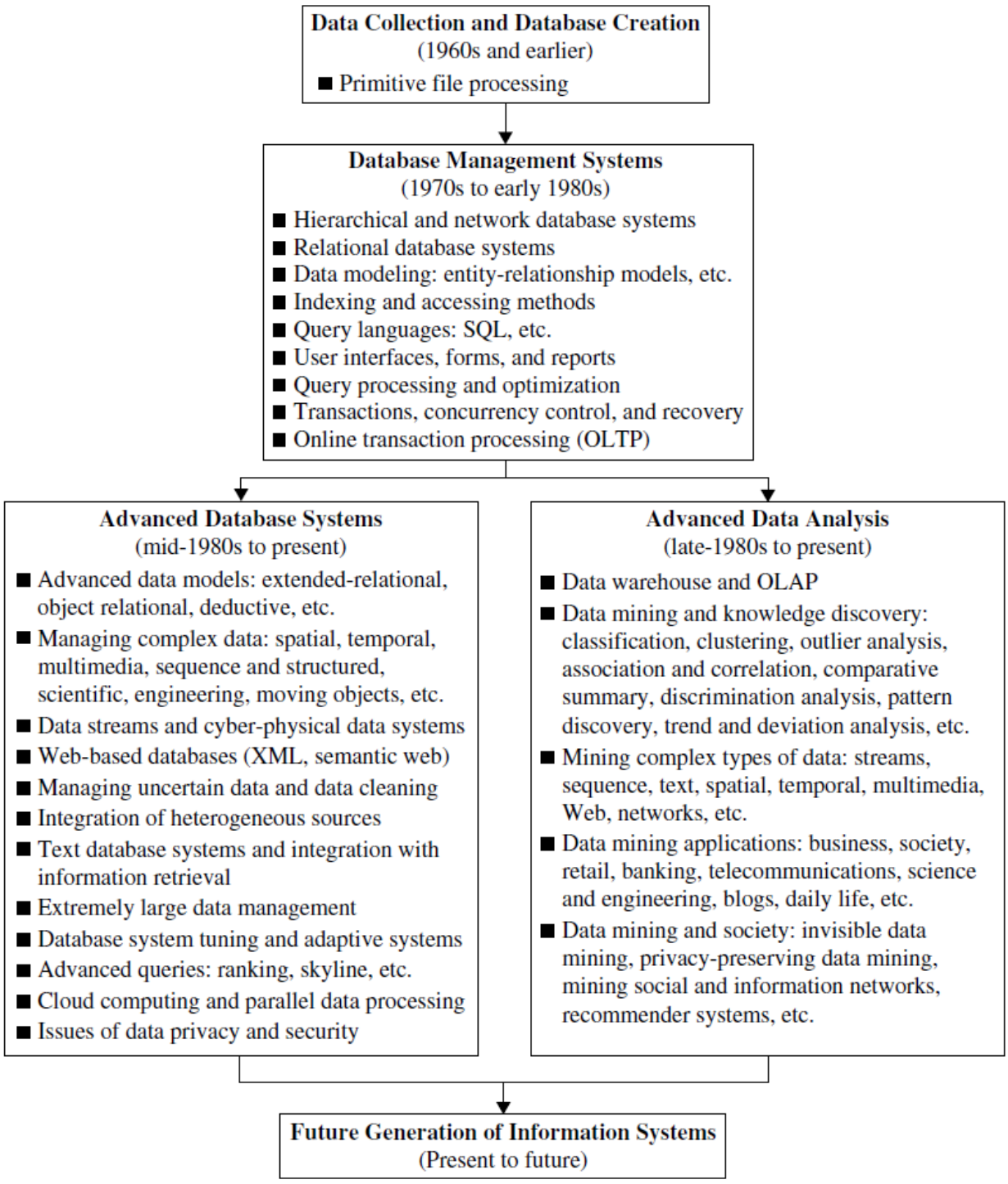
Търсачката (напр. Google) получава стотици милиони заявки всеки ден. Всяка заявка може да се разглежда като транзакция, при която потребителят описва своята нужда от информация. Какви нови и полезни знания може да научи една търсачка от такава огромна колекция от заявки, събрани от потребителите с течение на времето? Интересното е, че някои модели, открити в потребителските заявки за търсене, могат да разкрият безценни знания, които не могат да бъдат получени само чрез четене на отделни данни. Например Грипните тенденции на Google използват конкретни думи за търсене като индикатори за грипна активност. Той установи тясна връзка между броя на хората, които търсят информация, свързана с грипа, и броя на хората, които действително имат симптоми на грип. Появява се модел, когато се обобщят всички заявки за търсене, свързани с грип. Използвайки обобщени данни от търсенето с Google, Flu Trends може да оцени грипната активност до две седмици по-бързо от традиционните системи. Този пример показва как извличането на данни може да превърне голяма колекция от данни в знания, които могат да помогнат за посрещане на настоящо глобално предизвикателство.

Извличането на данни като еволюция на информационните технологии

Извличането на данни може да се разглежда като резултат от естествената еволюция на информационните технологии.

Индустрията за управление на бази данни и данни се разви в разработването на няколко критични функционалности (Фигура 1.1): събиране на данни и създаване на база данни, управление на данни (включително съхранение и извличане на данни и обработка на транзакции в базата данни) и разширен анализ на данни (включващ съхранение на данни и извличане на данни). Ранното развитие на механизмите за събиране на данни и създаване на база данни послужи като предпоставка за по-късното разработване на ефективни механизми за съхранение и извличане на данни, както и обработка на заявки и транзакции. В днешно време множество системи за бази данни предлагат обработка на заявки и транзакции като обичайна практика. Разширеният анализ на данни естествено се превърна в следващата стъпка.

От 60-те години на миналия век базата данни и информационните технологии се развиват системно от примитивни системи за обработка на файлове до сложни и мощни системи за бази данни.



Фиг. 1.1. Еволюцията на технологията на системата за бази данни.

Изследванията и разработките в системите за бази данни от 1970 г. напредват от ранните йерархични и мрежови системи за бази данни към системи за релационни бази данни (където данните се съхраняват в структури на релационни таблици), инструменти за моделиране на данни и методи за индексиране и достъп. Освен това потребителите получиха удобен и гъвкав достъп до данни чрез езици за заявки, потребителски интерфейси, оптимизация на заявки и управление на транзакции. Ефективните методи за онлайн обработка на транзакции (OLTP), при които заявката се разглежда като транзакция само за четене, допринесоха значително за еволюцията и широкото приемане на релационната технология като основен инструмент за ефективно съхранение, извличане и управление на големи количества данни.

След създаването на системи за управление на бази данни, технологията за бази данни се насочи към разработването на усъвършенствани системи за бази данни, съхранение на данни и извличане на данни за усъвършенстван анализ на данни и уеб базирани бази данни. Усъвършенстваните системи за бази данни, например, са резултат от подема на изследванията от средата на 80-те години нататък.

Тези системи включват нови и мощни модели на данни като разширено-релационни, обектно-ориентирани, обектно-релационни и дедуктивни модели. Приложно-ориентираните системи за бази данни процъфтяват, включително пространствени, времеви, мултимедийни, активни, поточни и сензорни, научни и инженерни бази данни, бази данни и офис информационни бази. Въпросите, свързани с разпространението, диверсификацията и споделянето на данни, са проучени широко.

Разширеният анализ на данни се появи от края на 80-те години на миналия век. Стабилният и ослепителен напредък на компютърната хардуерна технология през последните три десетилетия доведе до големи доставки на мощни и достъпни компютри, оборудване за събиране на данни и носители за съхранение. Тази технология осигурява голям тласък на базата данни и информационната индустрия и позволява огромен брой бази данни и информационни хранилища да бъдат достъпни за управление на транзакции, извличане на информация и анализ на данни. Данните вече могат да се съхраняват в много различни видове бази данни и информационни хранилища.

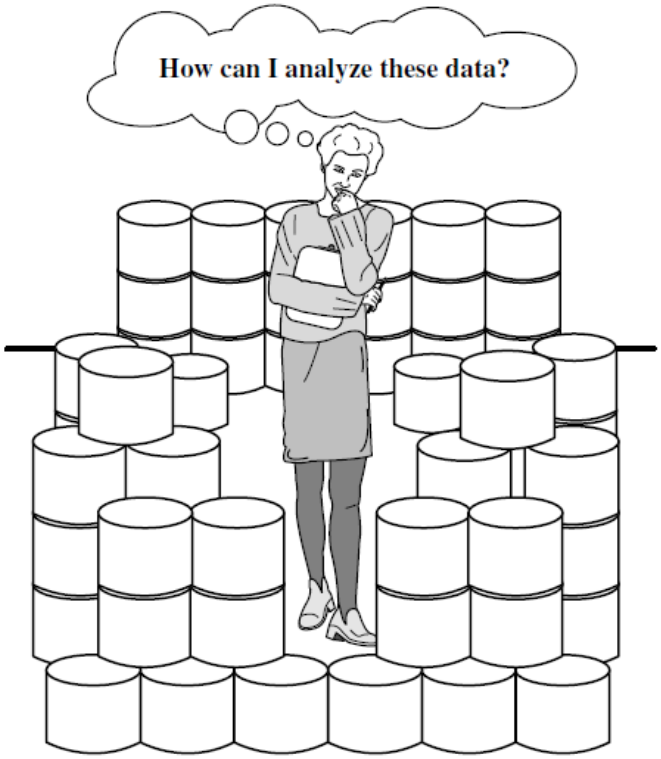
Една нововъзникваща архитектура за хранилище на данни е хранилището на данни.

Склад за данни е хранилище на множество разнородни източници на данни, организирани по унифицирана схема на едно място, за да се улесни вземането на управленски решения. Технологията за съхранение на данни включва почистване на данни, интегриране на данни и онлайн аналитична обработка (OLAP) – тоест техники за анализ с функционалности като обобщаване, консолидиране и агрегиране, както и възможност за преглед на информация от различни ъгли. Въпреки че OLAP инструментите поддържат многоизмерен анализ и вземане на решения, за задълбочен анализ са необходими допълнителни инструменти за анализ на данни - например инструменти за извличане на данни, които предоставят класификация на данни, групиране, откриване на отклонения/аномалии и характеризиране на промените в данните във времето.

Натрупани са огромни обеми данни извън базите данни и хранилищата за данни.

През 90-те години започват да се появяват World Wide Web и уеб-базирани бази данни (напр. XML бази данни). Интернет базирани глобални информационни бази, като WWW и различни видове взаимосвързани, разнородни бази данни, се появиха и играят жизненоважна роля в информационната индустрия. Ефективният и ефикасен анализ на данни от такива различни форми на данни чрез интегриране на технологии за извличане на информация, извличане на данни и анализ на информационни мрежи е предизвикателна задача.

В обобщение, изобилието от данни, съчетано с необходимостта от мощни инструменти за анализ на данни, е описано като богата на данни, но бедна на информация ситуация (Фигура 1.2).



Фиг. 1.2. Светът е богат на данни, но беден на информация

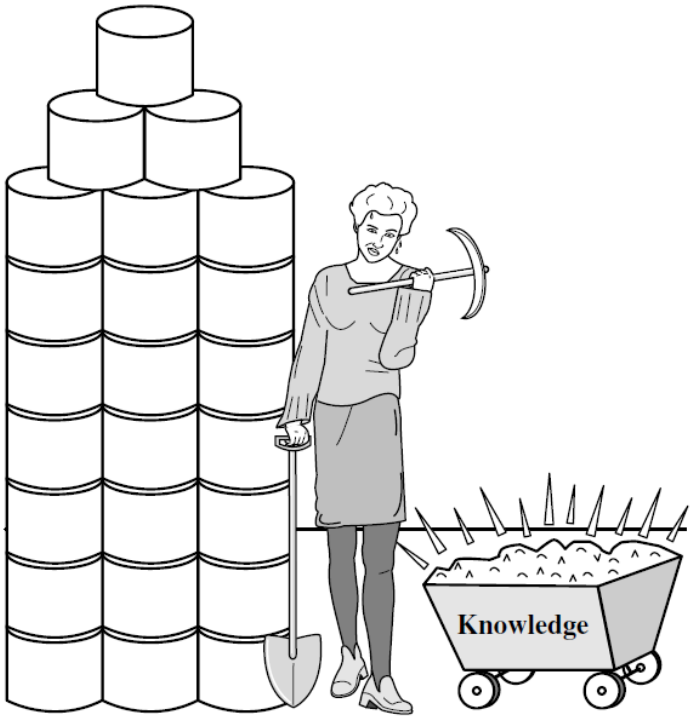
Бързо нарастващото, огромно количество данни, събрани и съхранявани в големи и многобройни хранилища на данни, далеч надхвърли човешката ни способност за разбиране без мощни инструменти. В резултат на това данните, събрани в големи хранилища на данни, се превръщат в „гробници с данни“ – архиви с данни, които рядко се посещават. Следователно важните решения често се вземат въз основа не на богатите на информация данни, съхранявани в хранилищата на данни, а по-скоро на интуицията на вземащия решения, просто защото този, който взема решения, не разполага с инструментите за извличане на ценните знания, вградени в огромните количества данни. Бяха положени усилия за разработване на експертна система и технологии, базирани на знания, които обикновено разчитат на потребители или експерти в областта да въвеждат ръчно знания в базите знания.

За съжаление обаче процедурата за ръчно въвеждане на знания е предразположена към пристрастия и грешки и е изключително скъпа и отнема време. Разширяващата се пропаст между данните и информацията изисква

систематично разработване на инструменти за извличане на данни, които могат да превърнат гробниците на данни в „златни късчета“ знания.

Какво е извличане на знания от данни?

Не е изненада, че извличането на данни, като наистина интердисциплинарен предмет, може да се дефинира по много различни начини. Дори терминът извличане на данни всъщност не представя всички основни компоненти в картината. За да се отнасяме до добива на злато от скали или пясък, казваме добив на злато вместо добив на скали или пясък. Аналогично, извличането на данни трябваше да бъде по-подходящо наречено „извличане на знания от данни“, което за съжаление е малко дълго. Въпреки това, в по-кратък план извличането на знания може да не отразява акцента върху извличането на големи количества данни. Независимо от това, добивът е ярък термин, характеризиращ процеса, който намира малък набор от скъпоценни късчета от голямо количество суровина (Фигура 1.3). По този начин такова погрешно наименование, носещо едновременно „данни“ и „копаене“, стана популярен избор. В допълнение,



Фиг. 1.3. Търсене на знания (модели) в данните.

Много хора разглеждат извличането на данни като синоним на друг популярно използван термин, откриване на знания от данни или KDD, докато други разглеждат извличането на данни като просто съществена стъпка в процеса на откриване на знания.

Процесът на откриване на знания е показан на фигура 1.4 като итеративна последователност от следните стъпки:

1. Филтриране на данни (за премахване на шум и непоследователни данни).

2. Интегриране на данни (където могат да се комбинират множество източници на данни).

3. Избор на данни (където данните, свързани със задачата за анализ, се извличат от базата данни).

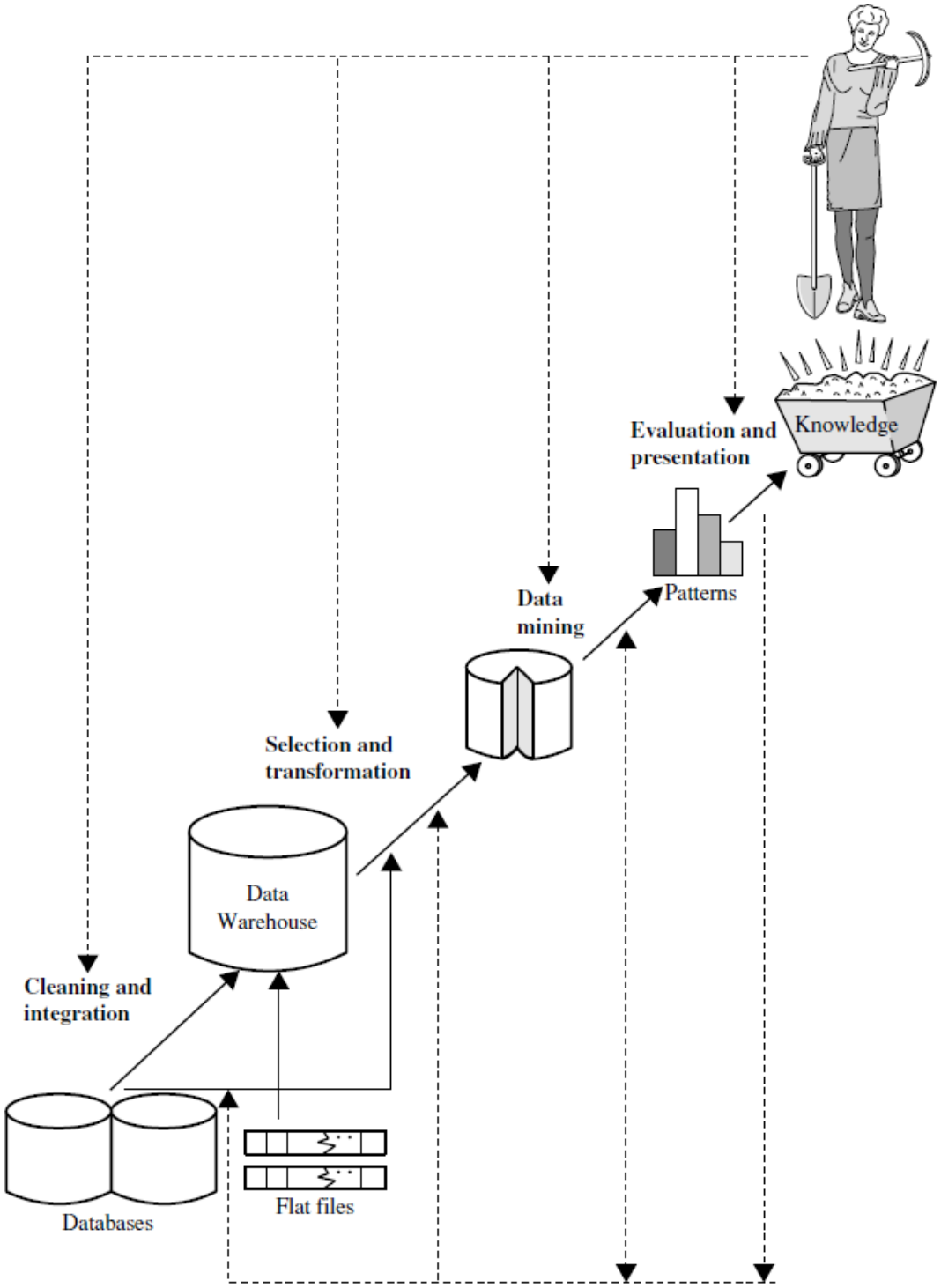
4. Трансформация на данни (където данните се трансформират и консолидират във форми, подходящи за копаене чрез извършване на операции за обобщаване или агрегиране).

5. Извличане на данни (съществен процес, при който се прилагат интелигентни методи за извличане на модели на данни).

6. Оценка на модела (за идентифициране на наистина интересните модели, представящи знания въз основа на мерки за интерес).

7. Представяне на знания (където техниките за визуализация и представяне на знания се използват за представяне на извлечени знания на потребителите).

Стъпки от 1 до 4 са различни форми на предварителна обработка на данни, при които данните се подготвят за копаене. Стъпката за извличане на данни може да взаимодейства с потребителя или базата от знания. Интересните модели се представят на потребителя и могат да се съхраняват като нови знания в базата знания.



Фиг. 1.4. Извличането на данни като стъпка в процеса на откриване на знания.

Предходният изглед показва извличането на данни като една стъпка в процеса на откриване на знания, макар и съществена, защото разкрива скрити модели за оценка. Въпреки това, в индустрията, в медиите и в изследователската среда терминът извличане на данни често се използва за обозначаване на целия процес на откриване на знания (може би защото терминът е по-кратък от откриване на знания от данни). Ето защо ние приемаме широк поглед върху функционалността за извличане на данни: Извличането на данни е процес на откриване на интересни модели и знания от големи количества данни. Източниците на данни могат да включват бази данни, хранилища на данни, уеб, други информационни хранилища или данни, които се предават динамично в системата.

Какви видове данни могат да бъдат изследвани?

Като обща технология извличането на данни може да се приложи към всякакъв вид данни, стига данните да са значими за целево приложение.

Най-основните форми на данни за приложения за копаене са данни от бази данни, данни от хранилище на данни и транзакционни данни.

Концепциите и техниките, представени в тази книга, се фокусират върху такива данни. Извличането на данни може да се приложи и към други форми на данни (напр. потоци от данни, подредени/последователни данни, графични или мрежови данни, пространствени данни, текстови данни, мултимедийни данни и WWW).

Извличането на данни със сигурност ще продължи да обхваща и новите типове данни, когато се появят.