БАЗИ ОТ ДАННИ – ХРАНИЛИЩА НА ГОЛЕМИ ДАННИ

Системата от бази от данни, наричана още система за управление на база данни (СУБД), се състои от колекция от взаимосвързани данни, известни като база данни, и набор от софтуерни програми за управление и достъп до данните. Софтуерните програми предоставят механизми за дефиниране на структури на бази данни и съхранение на данни; за определяне и управление на едновременен, споделен или разпределен достъп до данни; и за осигуряване на съгласуваност и сигурност на съхраняваната информация въпреки системни сривове или опити за неоторизиран достъп.

Релационна база данни е колекция от таблици, на всяка от които е присвоено уникално име. Всяка таблица се състои от набор от атрибути (колони или полета) и обикновено съхранява голям набор от кортежи (записи или редове). Всеки кортеж в релационна таблица представлява обект, идентифициран от уникален ключ и описан от набор от стойности на атрибути. Семантичен модел на данни, като например модел на данни на обект-връзка (ER), често се конструира за релационни бази данни. Един ER модел на данни представя базата данни като набор от обекти и техните взаимоотношения.

Пример 1.2 Релационна база данни за AllElectronics.

Фиктивният магазин AllElectronics се използва за илюстриране на концепции в тази книга. Компанията е описана от следните таблици на релации: клиент, артикул, служител и клон. Описаните тук таблици са показани на Фигура 1.5. (таблиците се наричат още схема или релация.)

Отношението клиент се състои от набор от атрибути, описващи информацията за клиента, включително уникален идентификационен номер на клиента (cust_ID), име на клиента, адрес, възраст, професия, годишен доход, кредитна информация и категория.

По същия начин, всеки от отношенията елемент, служител и клон се състои от набор от атрибути, описващи свойствата на тези обекти.

Таблиците могат да се използват и за представяне на връзките между или между множество обекти. В нашия пример те включват покупки (клиентът купува артикули, създава транзакция за продажба, управлявана от служител), продадени артикули (изброява артикулите, продадени в дадена транзакция) и работи в (служителят работи в клон на AllElectronics).

customer	*(cust_ID, name, address, age, occupation, annual_income, credit_information, category, . . .)*
item	*(item_ID, brand, category, type, price, place_made, supplier, cost, . . .)*
employee	*(empl_ID, name, category, group, salary, commission, . . .)*
branch	*(branch_ID, name, address, . . .)*
purchases	*(trans_ID, cust_ID, empl_ID, date, time, method_paid, amount)*
items_sold	*(trans_ID, item_ID, qty)*
works_at	*(empl_ID, branch_ID)*

Фиг. 1.5. Релационна схема за релационна база данни AllElectronics.

Релационните данни могат да бъдат достъпни чрез заявки към база данни, написани на език за релационни заявки (напр. SQL) или с помощта на графични потребителски интерфейси. Дадена заявка се трансформира в набор от релационни операции, като свързване, селекция и проекция, и след това се оптимизира за ефективна обработка. Заявката позволява извличане на определени подмножества от данните. Да предположим, че вашата работа е да анализирате данните на AllElectronics. Чрез използването на релационни заявки можете да питате неща като „Покажете ми списък с всички артикули, които са били продадени през последното тримесечие“.

Релационните езици също използват агрегатни функции като sum, avg (средно), count, max (максимум) и min (минимум). Използването на агрегати ви позволява да попитате: „Покажете ми общите продажби за последния месец, групирани по клонове“ или „Колко транзакции за продажба са извършени през месец декември?“ или „Кой търговец имаше най-високи продажби?“

Когато копаем релационни бази данни, можем да отидем по-далеч, като търсим тенденции или модели на данни. Например, системите за извличане на

данни могат да анализират данни за клиенти, за да предвидят кредитния риск на нови клиенти въз основа на техния доход, възраст и предишна кредитна информация. Системите за извличане на данни също могат да открият отклонения - т.е. артикули с продажби, които са далеч от очакваните в сравнение с предходната година. След това такива отклонения могат да бъдат допълнително изследвани. Например извличането на данни може да открие, че е имало промяна в опаковката на артикул или значително увеличение на цената.

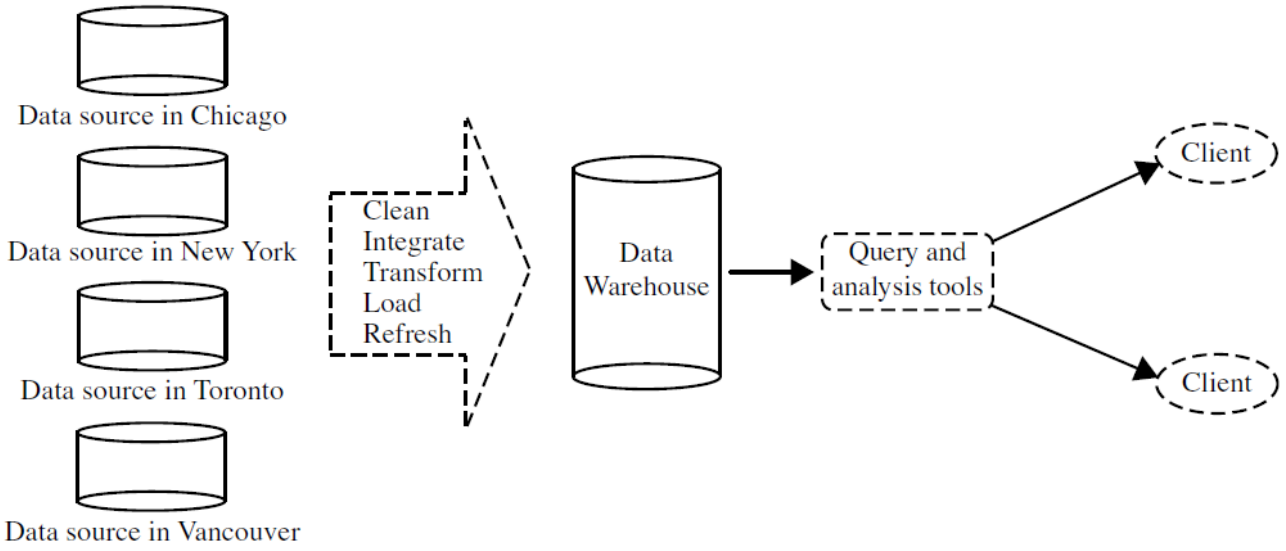
Релационните бази данни са едно от най-често достъпните и най-богатите хранилища на информация и по този начин те са основна форма на данни в изучаването на извличането на данни.

Хранилища за големи данни (Big Data Warehouses)

Да предположим, че AllElectronics е успешна международна компания с клонове по целия свят. Всеки клон има свой собствен набор от бази данни. Президентът на AllElectronics ви помоли да предоставите анализ на продажбите на компанията по вид артикул за клон за третото тримесечие. Това е трудна задача, особено след като съответните данни са разпределени в няколко бази данни, физически разположени на множество сайтове.

Ако AllElectronics имаше склад за данни, тази задача би била лесна. Складът за данни е хранилище на информация, събрана от множество източници, съхранявана под унифицирана схема и обикновено пребиваваща на един сайт. Складовете за данни се изграждат чрез процес на почистване на данни, интегриране на данни, трансформация на данни, зареждане на данни и периодично опресняване на данни. Този процес е разгледан в глави 3 и 4.

Фигура 1.6 показва типичната рамка за изграждане и използване на хранилище за данни за AllElectronics.



Фиг. 1.6. Типична рамка на хранилище за данни за AllElectronics.

За да се улесни вземането на решения, данните в хранилището на данни са организирани около основни субекти (напр. клиент, артикул, доставчик и дейност). Данните се съхраняват, за да предоставят информация от историческа гледна точка, като например през последните 6 до 12 месеца, и обикновено са обобщени. Например, вместо да съхранява подробностите за всяка продажбена транзакция, хранилището на данни може да съхранява обобщение на транзакциите по тип артикул за всеки магазин или, обобщени на по-високо ниво, за всеки регион на продажби.

Складът за данни обикновено се моделира от многомерна структура от данни, наречена куб с данни, в която всяко измерение съответства на атрибут или набор от атрибути в схемата и всяка клетка съхранява стойността на някаква агрегатна мярка, като например брой или сума (сума на продажбите).

Кубът с данни предоставя многоизмерен изглед на данните и позволява предварително изчисление и бърз достъп до обобщени данни.

Пример 1.3 Куб с данни за AllElectronics.

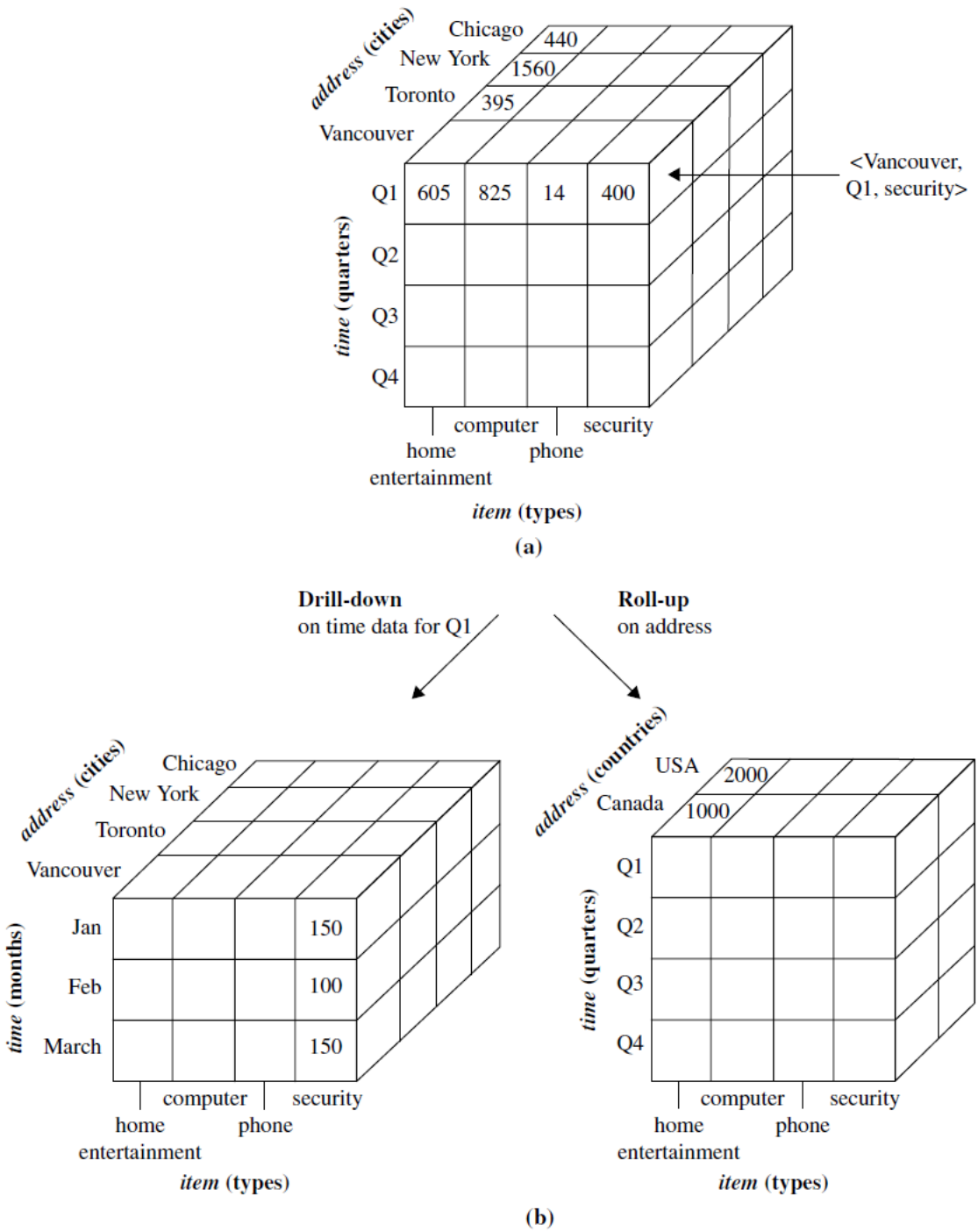
Куб с данни за обобщени данни за продажбите на AllElectronics е представен на Фигура 1.7(а). Кубът има три измерения: адрес (със стойности на град Чикаго, Ню Йорк, Торонто, Ванкувър), време (със стойности на тримесечие

Q1, Q2, Q3, Q4) и елемент (със стойности на типа на елемента домашно забавление, компютър, телефон, сигурност) . Общата стойност, съхранявана във всяка клетка на куба, е сумата на продажбите (в хиляди). Например общите продажби за първото тримесечие, Q1, за артикули, свързани със системи за сигурност във Ванкувър, са $400 000, както се съхраняват в клетка (Ванкувър, Q1, сигурност).

Допълнителни кубове могат да се използват за съхраняване на сборни суми за всяко измерение, съответстващи на сборните стойности, получени с помощта на различни SQL групи по (напр. общата сума на продажбите за град и тримесечие, или за град и артикул, или за тримесечие и артикул, или за всяко отделно измерение). Чрез предоставяне на многоизмерни изгледи на данни и предварително изчисление на обобщени данни, системите за съхранение на данни могат да осигурят присъща поддръжка за OLAP. Операциите за онлайн аналитична обработка използват основни знания по отношение на домейна на данните, които се изследват, за да позволят представянето на данни на различни нива на абстракция.

Такива операции се приспособяват към различни потребителски гледни точки. Примери за OLAP операции включват drill-down и roll-up, които позволяват на потребителя да преглежда данните с различни степени на обобщаване, както е показано на фигура 1.7 (b). Например, можем да разгледаме данните за продажбите, обобщени по тримесечия, за да видим данните, обобщени по месеци. По същия начин можем да обобщим данните за продажбите, обобщени по град, за да прегледаме данните, обобщени по държави.

Въпреки че инструментите за съхранение на данни помагат в подкрепа на анализа на данни, често са необходими допълнителни инструменти за извличане на данни за задълбочен анализ. Многомерното извличане на данни (наричано също проучвателно многоизмерно извличане на данни) извършва извличане на данни в многоизмерно пространство в OLAP стил. Това означава, че позволява изследването на множество комбинации от измерения на различни нива на детайлност в извличането на данни и по този начин има по-голям потенциал за откриване на интересни модели, представящи знания.



Фиг.1.7. Многоизмерен куб с данни, използван за описание на данни, (a) показване на обобщени данни за AllElectronics и (b) показване на обобщени данни, получени в резултат на операциите за разбивка надолу и събиране на куба в (a). За по-добра четливост са показани само някои от стойностите на кубичните клетки.

Транзакционни данни

Като цяло, всеки запис в транзакционна база данни улавя транзакция, като покупка на клиент, резервация за полет или щраквания на потребител върху уеб страница. Транзакцията обикновено включва уникален идентификационен номер на транзакция (trans ID) и списък на елементите, съставляващи транзакцията, като артикулите, закупени в транзакцията. Транзакционната база данни може да има допълнителни таблици, които съдържат друга информация, свързана с транзакциите, като описание на артикул, информация за продавача или клона и т.н.

Пример 1.4 Транзакционна база данни за AllElectronics.

Транзакциите могат да се съхраняват в таблица, с един запис на транзакция. Фрагмент от транзакционна база данни за AllElectronics е показан на Фигура 1.8. От гледна точка на релационната база данни, таблицата за продажби на фигурата е вложена релация, тъй като списъкът с атрибути с идентификатори на артикули съдържа набор от артикули.

Тъй като повечето системи за релационни бази данни не поддържат вложени релационни структури, транзакционната база данни обикновено се съхранява в плосък файл във формат, подобен на таблицата на Фигура 1.8, или се разгръща в стандартна релация във формат, подобен на таблицата за продадени артикули на Фигура 1.5.

Като анализатор на AllElectronics може да попитате „Кои артикули се продават добре заедно?“ Този вид анализ на данни за пазарната кошница ще ви позволи да групирате групи от артикули заедно като стратегия за увеличаване на продажбите. Например, имайки предвид, че принтерите обикновено се купуват заедно с компютри, бихте могли да предложите определени принтери с голяма отстъпка (или дори безплатно) на клиенти, които купуват избрани компютри, с надеждата да продадете повече компютри (които често са по-скъпи от принтери). Традиционната система от бази данни не е в състояние да извърши анализ на данни за пазарната кошница. За щастие извличането на данни върху транзакционни данни може да направи това чрез копаене на чести набори от артикули, тоест набори от артикули, които често се продават заедно.

Други видове данни

Освен данни от релационни бази данни, данни от хранилища на данни и данни за транзакции, има много други видове данни, които имат разнообразни форми и структури и доста различни семантични значения. Такива видове данни могат да се видят в много приложения: свързани с времето или последователни данни (напр. исторически записи, борсови данни и времеви редове и данни за биологична последователност), потоци от данни (напр. данни за видеонаблюдение и сензори, които са непрекъснато предавани), пространствени данни (напр. карти), данни за инженерен дизайн (напр. дизайн на сгради, системни компоненти или интегрални схеми), хипертекст и мултимедийни данни (включително текст, изображения, видео и аудио данни), графики и мрежови данни (напр. социални и информационни мрежи) и мрежата (огромно, широко разпространено хранилище на информация, предоставено от интернет).

Различни видове знания могат да бъдат извлечени от тези видове данни. Тук изброяваме само няколко. По отношение на временните данни, например, можем да извличаме банкови данни за променящи се тенденции, което може да помогне при планирането на банковите касиери според обема на клиентския трафик. Данните от борсата могат да бъдат извлечени, за да се разкрият тенденции, които могат да ви помогнат да планирате инвестиционни стратегии (напр. най-доброто време за закупуване на акции на AllElectronics). Можем да копаем потоци от данни в компютърна мрежа, за да открием проникванията въз основа на аномалията на потоците от съобщения, които могат да бъдат открити чрез клъстериране, динамично изграждане на модели на потоци или чрез сравняване на текущите чести модели с тези в предишно време.

С пространствени данни можем да търсим модели, които описват промените в нивата на бедност в метрополисите въз основа на разстоянията на града от главните магистрали. Връзките между набор от пространствени обекти могат да бъдат изследвани, за да се открие кои подмножества от обекти са пространствено автокорелирани или свързани. Чрез извличане на текстови данни, като например литература за извличане на данни от последните десет години, можем да идентифицираме развитието на горещи теми в тази област. Чрез извличане на потребителски коментари за продукти (които често се

изпращат като кратки текстови съобщения), можем да оценим настроенията на клиентите и да разберем колко добре даден продукт е възприет от пазара. От мултимедийни данни можем да извличаме изображения, за да идентифицираме обекти и да ги класифицираме чрез присвояване на семантични етикети или тагове. Чрез извличане на видео данни от хокеен мач можем да открием видео последователности, съответстващи на голове.

Важно е да имате предвид, че в много приложения присъстват множество типове данни. Например при уеб копаене често има текстови данни и мултимедийни данни (напр. снимки и видеоклипове) на уеб страници, графични данни като уеб графики и картографски данни на някои уеб сайтове. В биоинформатиката геномни последователности, биологични мрежи и 3-D пространствени структури на геноми могат да съществуват съвместно за определени биологични обекти. Извличането на множество източници на данни на сложни данни често води до ползотворни открития поради взаимното подобряване и консолидиране на такива множество източници. От друга страна, това също е предизвикателство поради трудностите при почистването и интегрирането на данни, както и сложните взаимодействия между множеството източници на такива данни.

Въпреки че такива данни изискват сложни съоръжения за ефективно съхранение, извличане и актуализиране, те също така осигуряват благоприятна почва и повдигат предизвикателни проблеми с изследванията и прилагането на извличането на данни. Извличането на данни върху такива данни е тема за напреднали.

Какви видове модели могат да се изследват?

Наблюдавахме различни видове хранилища за данни и информация, върху които може да се извършва извличане на данни. Нека сега разгледаме видовете модели, които могат да бъдат добивани.

Има редица функции за извличане на данни. Те включват:

- характеризиране и дискриминация;
- извличането на модели,
- асоциации и корелации;
- класификация и регресия;

Проект BG05M2OP001-2.016-0003 „Модернизация на Национален военен университет "Васил Левски"- гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, финансиран от ниски съюз чрез Европейските структурни и инвестиционни фондове.

- клъстерен анализ;
- анализ на отклоненията.

Функционалностите за извличане на данни се използват за определяне на видовете модели, които да бъдат открити в задачите за извличане на данни. Най-общо такива задачи могат да бъдат класифицирани в две категории: описателни и прогнозни. Описателните задачи за копаене характеризират свойствата на данните в целеви набор от данни. Задачите за предсказуем добив извършват индукция върху текущите данни, за да направят прогнози. Функционалностите за извличане на данни и видовете модели, които могат да открият, са описани по-долу. Интересните модели представляват знание.

Описание на клас/концепция: Характеризиране и дискриминация

Записите на данни могат да бъдат свързани с класове или концепции. Например в магазина AllElectronics класовете артикули за продажба включват компютри и принтери, а концепциите за клиенти включват bigSpenders и budgetSpenders. Може да бъде полезно да се опишат отделни класове и концепции в обобщени, стегнати и все пак точни термини. Такива описания на клас или концепция се наричат описания на клас/концепция. Тези описания могат да бъдат получени с помощта на (1) характеризиране на данни, чрез обобщаване на данните от изследвания клас (често наричан целеви клас) в общи термини, или (2) разграничаване на данни, чрез сравнение на целевия клас с един или набор на сравнителни класове (често наричани контрастиращи класове) или (3) както характеризиране на данни, така и дискриминация.

Характеризиране на данните е обобщение на общите характеристики или характеристики на целеви клас данни. Данните, съответстващи на посочения от потребителя клас, обикновено се събират чрез заявка. Например, за да се проучат характеристиките на софтуерни продукти с продажби, които са се увеличили с 10% през предходната година, данните, свързани с такива продукти, могат да бъдат събрани чрез изпълнение на SQL заявка в базата данни за продажби.

Има няколко метода за ефективно обобщаване и характеризиране на данни.

Базираната на куб данни OLAP сборна операция (Раздел 1.3.2) може да се използва за извършване на контролирано от потребителя обобщаване на данни по определено измерение. Индукционна техника, ориентирана към атрибути, може да се използва за извършване на генерализиране и характеризиране на данни без потребителско взаимодействие стъпка по стъпка.

Резултатът от характеризирането на данните може да бъде представен в различни форми. Примерите включват кръгови диаграми, стълбовидни диаграми, криви, многоизмерни кубове с данни и многоизмерни таблици, включително кръстосани таблици. Получените описания могат също да бъдат представени като обобщени отношения или под формата на правила (наречени характеристични правила).

Пример 1.5 Характеризиране на данните.

Мениджърът за връзки с клиенти в AllElectronics може да поръча следната задача за извличане на данни: Обобщете характеристиките на клиентите, които харчат повече от $5000 годишно в AllElectronics. Резултатът е общ профил на тези клиенти, като например, че са на възраст от 40 до 50 години, работят и имат отличен кредитен рейтинг. Системата за извличане на данни трябва да позволи на мениджъра за взаимоотношения с клиенти да се задълбочи във всяко измерение, като например професия, за да види тези клиенти според вида им на работа.

Дискриминация на данни е сравнение на общите характеристики на обектите с данни от целевия клас спрямо общите характеристики на обекти от един или множество контрастни класове.

Целевият и контрастиращият клас могат да бъдат зададени от потребител и съответните обекти с данни могат да бъдат извлечени чрез заявки към база данни. Например, даден потребител може да поиска да сравни общите характеристики на софтуерни продукти с продажби, които са се увеличили с 10% миналата година, спрямо тези с продажби, които са намалели с поне 30% през същия период. Методите, използвани за дискриминация на данни, са подобни на тези, използвани за характеризиране на данни.

„Как се извеждат описанията на дискриминацията?“ Формите на представяне на изхода са подобни на тези за описания на характеристиките, въпреки че описанията на дискриминацията трябва да включват сравнителни мерки, които помагат да се разграничат целевите и контрастиращите класове. Описанията на дискриминацията, изразени под формата на правила, се наричат дискриминационни правила.

Пример 1.6 Дискриминация на данни.

Мениджърът за връзки с клиенти в AllElectronics може да поиска да сравни две групи клиенти – тези, които пазаруват компютърни продукти редовно (напр. повече от два пъти месечно) и тези, които рядко пазаруват такива продукти (напр. по-малко от три пъти годишно). Полученото описание предоставя общ сравнителен профил на тези клиенти, като например, че 80% от клиентите, които често купуват компютърни продукти, са на възраст между 20 и 40 години и имат висше образование, докато 60% от клиентите, които рядко купуват такива продукти, са възрастни или младежи и нямат висше образование. Задълбочаването на измерение като професия или добавянето на ново измерение като ниво на дохода може да помогне да се намерят още повече разграничителни характеристики между двата класа.

Модели, асоциации и корелации

Често срещани модели, както подсказва името, са модели, които се срещат често, повтарят се в данните.

Има много видове често срещани модели, включително често срещани набори от елементи, чести подпоследователности (известни също като последователни модели) и чести подструктури. Често срещаният набор от артикули обикновено се отнася до набор от артикули, които често се появяват заедно в набор от транзакционни данни – например мляко и хляб, които често се купуват заедно в магазините за хранителни стоки от много клиенти. Често срещана подпоследователност, като модела, че клиентите са склонни да закупят първо лаптоп, последван от цифров фотоапарат и след това карта с памет, е (чест) последователен модел. Една подструктура може да се отнася до различни структурни форми (напр. графики, дървета или решетки), които могат да се

комбинират с набори от елементи или подпоследователности. Ако субструктура се появява често, тя се нарича (често) структуриран модел.

Пример 1.7 Анализ на асоциацията.

Да предположим, че като маркетинг мениджър в AllElectronics искате да знаете кои елементи често се купуват заедно (т.е. в рамките на една и съща сделка).

Пример за такова правило, извлечено от транзакционната база данни на AllElectronics, е buys.X, "компютър"/)buys.X, "софтуер"/ [поддръжка D 1%, увереност D 50%], където X е променлива, представляваща клиент. Доверие или сигурност от 50% означава, че ако клиент купи компютър, има 50% шанс той да закупи и софтуер. Поддръжка от 1% означава, че 1% от всички анализирани транзакции показват, че компютърът и софтуерът са закупени заедно. Това правило за асоцииране включва един атрибут или предикат (т.е. купува), който се повтаря. Правилата за асоцииране, които съдържат един предикат, се наричат едномерни правила за асоцииране. Изпускайки предикатната нотация, правилото може да бъде написано просто като „компютърен софтуер [1%, 50%]“.

Да предположим вместо това, че ни е дадена релационната база данни AllElectronics, свързана с покупките.

Система за извличане на данни може да намери правила за асоцииране като

възраст(X, “20..29”) ^ доход (X, “40K..49K”) => купува (X, “лаптоп”)
[подкрепа = 2%, увереност = 60%].

Правилото показва, че от изследваните клиенти на AllElectronics 2% са на възраст от 20 до 29 години с доход от $40 000 до $49 000 и са закупили лаптоп (компютър) от AllElectronics. Има 60% вероятност клиент от тази възрастова и доходна група да закупи лаптоп. Имайте предвид, че това е асоциация, включваща повече от един атрибут или предикат (т.е. възраст, доход и покупки). Възприемайки терминологията, използвана в многомерните бази данни, където всеки атрибут се нарича измерение, горното правило може да се нарече правило за многомерно асоцииране.

Обикновено правилата за асоцииране се отхвърлят като безинтересни, ако не отговарят както на минимален праг на поддръжка, така и на минимален праг на доверие. Може да се извърши допълнителен анализ, за да се разкрият интересни статистически корелации между свързаните двойки атрибут-стойност.

*Често копаене на набор от елементи*е основна форма на често копаене на модели.

Извличането на последователни модели и извличането на структурирани модели се считат за теми за напреднали.

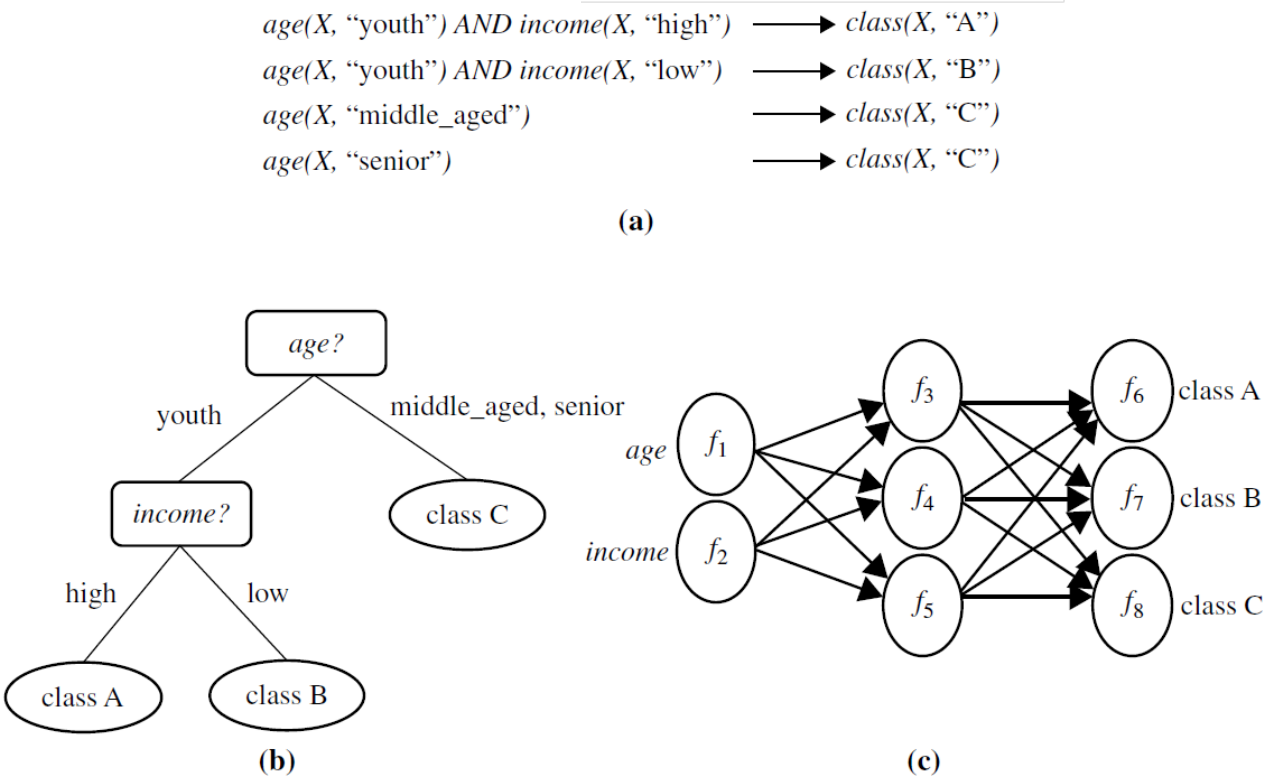
Класификация и регресия за прогнозен анализ

Класификация е процес на намиране на модел (или функция), който описва и разграничава класове данни или концепции. Моделът се извлича въз основа на анализ на набор от данни за обучение (т.е. обекти с данни, за които са известни етикетите на класа). Моделът се използва за прогнозиране на етикета на класа на обекти, за които етикетът на класа е неизвестен.

„Как се представя извлеченият модел?“ Изведеният модел може да бъде представен в различни форми, като правила за класификация (т.е. правила АКО-ТОГАВА), дървета на решенията, математически формули или невронни мрежи (Фигура 1.9).

Дървото на решенията е дървовидна структура, подобна на блок-схема, където всеки възел обозначава тест на стойност на атрибут, всеки клон представлява резултат от теста, а листата на дървото представляват класове или класови разпределения.

Дърветата на решенията могат лесно да бъдат преобразувани в правила за класификация. Невронната мрежа, когато се използва за класификация, обикновено е колекция от невроноподобни обработващи единици с претеглени връзки между единиците. Има много други методи за конструиране на класификационни модели, като наивна байесова класификация, опорни векторни машини и класификация на k-най-близкия съсед.



Фиг. 1.9. Класификационният модел може да бъде представен в различни форми:
(a) правила АКО-ТОГАВА, (b) дърво на решенията или (c) невронна мрежа.

Докато класификацията предвижда категорични (дискретни, неподредени) етикети, регресионните модели непрекъснато оценени функции. Това означава, че регресията се използва за прогнозиране на липсващи или неналични числени стойности на данни, а не (дискретни) етикети на класове. Терминът прогнозиране се отнася както за числово прогнозиране, така и за прогнозиране на етикет на клас. Регресионният анализ е статистическа методология, която най-често се използва за числено прогнозиране, въпреки че съществуват и други методи. Регресията също така обхваща идентифицирането на тенденциите на разпространение въз основа на наличните данни.

Може да се наложи класификацията и регресията да бъдат предшествани от анализ на релевантността, който се опитва да идентифицира атрибути, които са значително релевантни за процеса на класификация и регресия. Такива

атрибути ще бъдат избрани за процеса на класификация и регресия. След това други атрибути, които са неуместни, могат да бъдат изключени от разглеждане.

Пример 1.8 Класификация и регресия.

Да предположим, че като мениджър продажби на AllElectronics искате да класифицирате голям набор от артикули в магазина въз основа на три вида отговори на кампания за продажби: добър отговор, слаб отговор и липса на отговор. Искате да извлечете модел за всеки от тези три класа въз основа на описателните характеристики на артикулите, като цена, марка, място на производство, тип и категория. Получената класификация трябва максимално да разграничава всеки клас от останалите, представяйки организирана картина на набора от данни.

Да предположим, че получената класификация е изразена като дърво на решенията. Дървото на решенията, например, може да идентифицира цената като единствения фактор, който най-добре разграничава трите класа. Дървото може да разкрие, че в допълнение към цената, други характеристики, които помагат за по-нататъшното разграничаване на обектите от всеки клас един от друг, включват марка и място на производство.

Такова дърво на решенията може да ви помогне да разберете въздействието на дадена кампания за продажби и да проектирате по-ефективна кампания в бъдеще.

Да предположим вместо това, че вместо да предвиждате категорични етикети за отговор за всеки артикул от магазина, бихте искали да предвидите сумата на приходите, които всеки артикул ще генерира по време на предстояща продажба в AllElectronics, въз основа на предишни данни за продажби. Това е пример за регресионен анализ, тъй като конструираният регресионен модел ще предвиди непрекъсната функция (или подредена стойност).

Клъстерен анализ

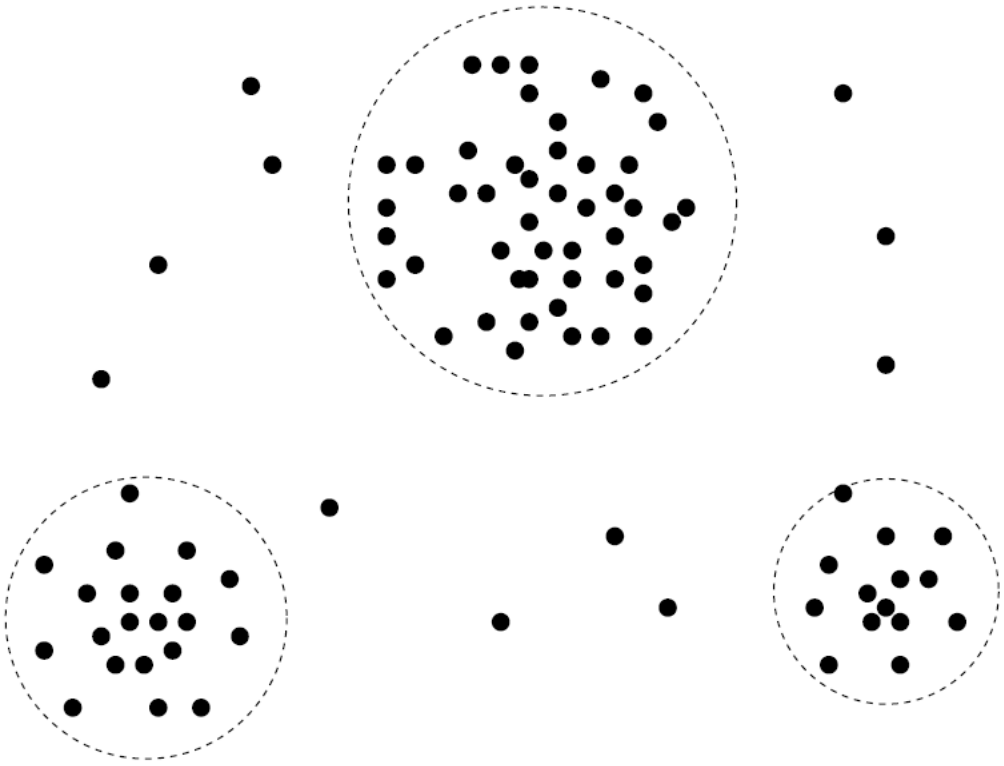
За разлика от класификацията и регресията, които анализират маркирани с класове (обучителни) набори от данни, клъстерирането анализира обекти с данни без да се консултира с етикети на класове. В много случаи данните, обозначени с клас, може просто да не съществуват в началото. Клъстерирането може да се

използва за генерираненаклас етикети за група от данни. Обектите са клъстерирани или групирани въз основа на принципа за максимизиране на вътрешнокласовото сходство и минимизиране на междукласовото сходство. Това означава, че клъстерите от обекти се формират така, че обектите в рамките на клъстера имат голямо сходство в сравнение един с друг, но са доста различни от обектите в други клъстери. Всеки така формиран клъстер може да се разглежда като клас от обекти, от които могат да се извличат правила. Групирането може също да улесни формирането на таксономия, тоест организирането на наблюденията в йерархия от класове, които групират подобни събития заедно.

Пример 1.9. Клъстерен анализ.

Клъстерният анализ може да се извърши върху клиентски данни на AllElectronics, за да се идентифицират хомогенни подпопулации от клиенти. Тези клъстери могат да представляват отделни целеви групи за маркетинг.

Фигура 1.10 показва 2-D диаграма на клиентите по отношение на местоположението на клиентите в даден град. Три групи точки от данни са очевидни.



Фиг.1.10 .Двуизмерна графика на клиентски данни по отношение на клиентските местоположения в град, показваща три клъстера от данни.

Анализ на отклоненията

Набор от данни може да съдържа обекти, които не отговарят на общото поведение или модел на данните. Тези обекти с данни са извънредни стойности. Много методи за извличане на данни отхвърлят отклоненията като шум или изключения. Въпреки това, в някои приложения (напр. откриване на измами) редките събития могат да бъдат по-интересни от по-редовните. Анализът на външни данни се нарича анализ на извънредни стойности или извличане на аномалии.

Излизащите стойности могат да бъдат открити с помощта на статистически тестове, които предполагат модел на разпределение или вероятност за данните, или с помощта на мерки за разстояние, където обекти, които са отдалечени от всеки друг клъстер, се считат за извънредни стойности. Вместо да използват статистически мерки или мерки за разстояние, методите, базирани на плътност, могат да идентифицират отклонения в локален регион, въпреки че те изглеждат нормални от изглед на глобално статистическо разпределение.

Пример 1.10. Анализ на отклонения.

Анализът на отклоненията може да разкрие измамно използване на кредитни карти чрез откриване на покупки на необичайно големи суми за даден номер на сметка в сравнение с обичайните такси, направени от същата сметка. Извънредните стойности могат също да бъдат открити по отношение на местоположенията и видовете покупки или честотата на покупките.

Всички модели ли са интересни?

Системата за извличане на данни има потенциала да генерира хиляди или дори милиони шаблони или правила.

Може да попитате: „Всички модели интересни ли са?“ Обикновено отговорът е не - само малка част от потенциално генерираните модели всъщност биха представлявали интерес за даден потребител.

Това повдига някои сериозни въпроси за извличането на данни. Може да се чудите: „Какво прави един модел интересен? Може ли система за извличане на данни да генерира всички интересни модели? Или може ли системата да генерира само интересните?“

За да отговоря на първия въпрос, моделът е интересен, ако е (1) лесно разбираем от хората, (2) валиден за нови или тестови данни с известна степен на сигурност, (3) потенциално полезен и (4) нов. Моделът също е интересен, ако потвърждава хипотеза, която потребителят иска да потвърди. Интересен модел представлява знание.

Съществуват няколко обективни мерки за интересност на модела. Те се основават на структурата на откритите модели и статистиката, която ги основава. Обективна мярка за правила за асоцииране от формата X)Y е поддръжката на правило, представляваща процента на транзакциите от база данни за транзакции, които даденото правило удовлетворява. Това се приема за вероятността PX [Y/, където X [Y показва, че транзакцията съдържа както X, така и Y, т.е. обединението на артикули X и Y. Друга обективна мярка за правилата за асоцииране е доверието, което оценява степента на на сигурността на откритата асоциация. Това се приема за условна вероятност P.YjX), тоест вероятността транзакция, съдържаща X, също да съдържа Y. По-формално, подкрепата и увереността се определят като:

$$\textit{поддържа}(X \Rightarrow Y) = P(XUY),$$
$$\textit{увереност}(X \Rightarrow Y) = P(Y \mid X).$$

Като цяло всяка мярка за интерес е свързана с праг, който може да се контролира от потребителя. Например правила, които не отговарят на праг на доверие от, да речем, 50%, могат да се считат за безинтересни. Правилата под прага вероятно отразяват шум, изключения или малцинствени случаи и вероятно са с по-малка стойност.

Други обективни мерки за интерес включват точност и покритие за правилата за класификация (АКО-ТОГАВА). Най-общо казано, точността ни казва процента на данните, които са правилно класифицирани от правило. Покритието е подобно на поддръжката, тъй като ни казва процента от данните, към които се прилага дадено правило. По отношение на разбираемостта, можем да използваме прости обективни мерки, които оценяват сложността или дължината в битове на извлечените модели.

Въпреки че обективните мерки помагат да се идентифицират интересни модели, те често са недостатъчни, освен ако не се комбинират със субективни мерки, които отразяват нуждите и интересите на конкретен потребител. Например моделите, описващи характеристиките на клиентите, които пазаруват често от AllElectronics, трябва да са интересни за маркетинг мениджъра, но може да са малко интересни за други анализатори, изучаващи същата база данни за модели върху представянето на служителите. Освен това много модели, които са интересни по обективни стандарти, могат да представляват здрав разум и следователно всъщност са безинтересни.

Мерки за субективна интересност се основават на вярванията на потребителите в данните. Тези мерки намират моделите за интересни, ако моделите са неочаквани (противоречащи на убежденията на потребителя) или предлагат стратегическа информация, въз основа на която потребителят може да действа. В последния случай такива модели се наричат подлежащи на действие. Например, модели като „голямо земетресение често следва група от малки земетресения“ може да са много приложими, ако потребителите могат да действат въз основа на информацията, за да спасят животи. Моделите, които се очакват, могат да бъдат интересни, ако потвърждават хипотеза, която потребителят желае да потвърди, или приличат на предчувствието на потребителя.

Вторият въпрос - „Може ли система за извличане на данни да генерира всички интересни модели?“ - се отнася до пълнотата на алгоритъм за извличане на данни. Често е нереалистично и неефективно системите за извличане на данни да генерират всички възможни модели. Вместо това трябва да се използват предоставени от потребителя ограничения и мерки за интерес, за да се фокусира търсенето.

За някои задачи за копаене, като асоцииране, това често е достатъчно, за да се гарантира пълнотата на алгоритъма. Извличането на правила за асоцииране е пример, при който използването на ограничения и мерки за интерес може да гарантира пълнотата на извличането.

И накрая, третият въпрос - „Може ли една система за извличане на данни да генерира само интересни модели?“ - е проблем за оптимизиране в извличането на данни. Много е желателно системите за извличане на данни да генерират само интересни модели. Това би било ефективно за потребителите и системите за извличане на данни, тъй като нито един от тях няма да трябва да търси в генерираните модели, за да идентифицира наистина интересните. В тази посока е постигнат напредък; подобна оптимизация обаче остава предизвикателство при извличането на данни.

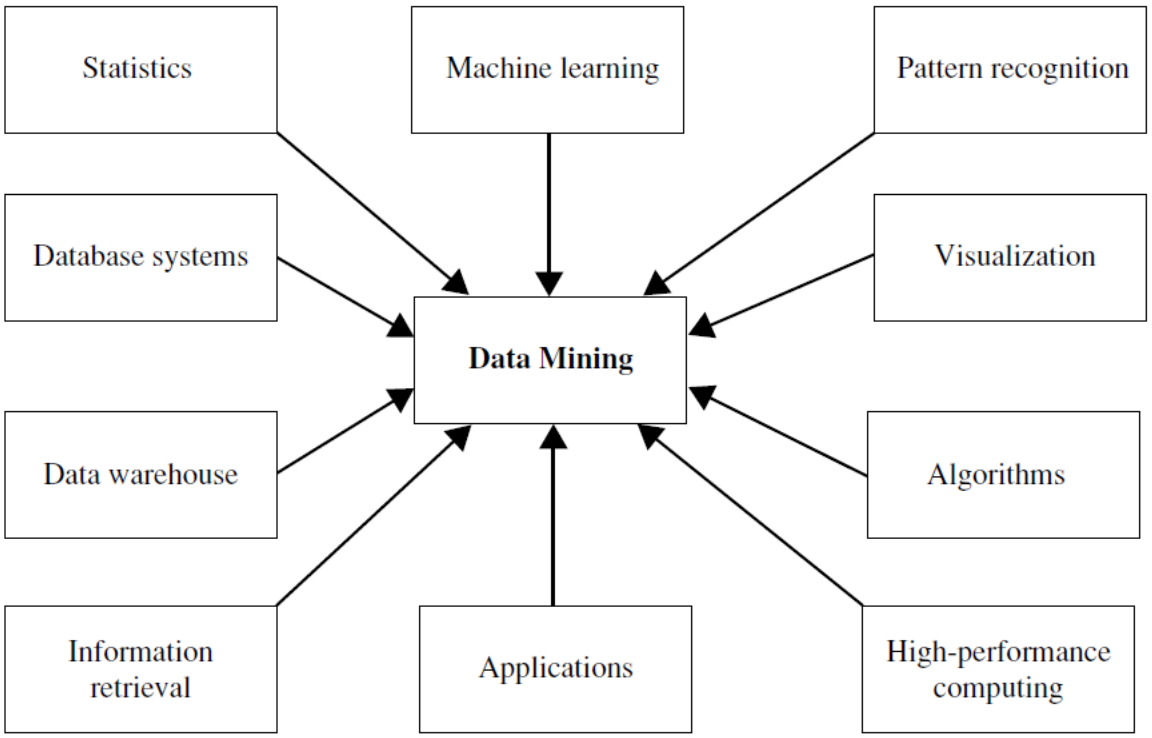
Мерките за интересност на моделите са от съществено значение за ефективното откриване на модели от целевите потребители. Такива мерки могат да се използват след стъпката за извличане на данни, за да се класират откритите модели според тяхната интересност, като се филтрират неинтересните.

По-важното е, че такива мерки могат да се използват за насочване и ограничаване на процеса на откриване, подобрявайки ефективността на търсенето чрез отрязване на подмножества от пространството на шаблона, които не отговарят на предварително определени ограничения за интерес.

Кои технологии се използват?

Като силно управляван от приложения домейн, извличането на данни включва много техники от други области като статистика, машинно обучение, разпознаване на модели, бази данни и системи за съхранение на данни, извличане на информация, визуализация, алгоритми, високопроизводителни изчисления и много приложни домейни (Фигура 1.11).

Интердисциплинарният характер на изследването и развитието на извличането на данни допринася значително за успеха на извличането на данни и неговите обширни приложения. В този раздел ние даваме примери за няколко дисциплини, които силно влияят върху развитието на методите за извличане на данни.



Фиг. 1.11. Извличането на данни приема техники от много области.

Статистика

Статистикаизучава събирането, анализа, тълкуването или обяснението и представянето на данни. Извличането на данни има присъща връзка със статистиката.

Статистическият модел е набор от математически функции, които описват поведението на обектите в целеви клас по отношение на случайни променливи и свързаните с тях вероятностни разпределения. Статистическите модели се използват широко за моделиране на данни и класове данни.

Например при задачи за извличане на данни като характеризиране и класификация на данни могат да бъдат изградени статистически модели на целеви класове. С други думи, такива статистически модели могат да бъдат резултат от задача за извличане на данни. Като алтернатива, задачите за извличане на данни могат да бъдат изградени върху статистически модели. Например, можем да използваме статистика, за да моделираме шум и стойности на липсващи данни. След това, когато извлича модели в голям набор от данни,

процесът на извличане на данни може да използва модела, за да помогне за идентифициране и обработка на шумни или липсващи стойности в данните.

Статистическите изследвания разработват инструменти за предвиждане и прогнозиране с помощта на данни и статистически модели. Статистическите методи могат да се използват за обобщаване или описание на колекция от данни.

Основните статистически описания на данните са въведени в глава 2.

Статистиката е полезна за извличане на различни модели от данни, както и за разбиране на основните механизми, генериращи и засягащи моделите. Инференциалната статистика (или прогнозната статистика) моделира данните по начин, който отчита случайността и несигурността в наблюденията и се използва за извеждане на изводи относно процеса или популацията, които се изследват.

Статистическите методи могат да се използват и за проверка на резултатите от извличане на данни. Например, след извличане на модел за класификация или прогнозиране, моделът трябва да бъде проверен чрез тестване на статистическа хипотеза. Тестът за статистическа хипотеза (понякога наричан анализ на потвърдителни данни) прави статистически решения, използвайки експериментални данни. Резултат се нарича статистически значим, ако е малко вероятно да е възникнал случайно. Ако моделът за класификация или прогнозиране е верен, тогава описателната статистика на модела увеличава надеждността на модела.

Прилагането на статистически методи в извличането на данни далеч не е тривиално. Често сериозно предизвикателство е как да се разшири статистически метод върху голям набор от данни. Много статистически методи имат висока сложност при изчисление. Когато такива методи се прилагат върху големи масиви от данни, които също са разпределени на множество логически или физически сайтове, алгоритмите трябва да бъдат внимателно проектирани и настроени, за да се намалят изчислителните разходи. Това предизвикателство става още по-трудно за онлайн приложения, като предложения за онлайн заявки в търсачките, където се изисква извличане на данни за непрекъснато обработване на бързи потоци от данни в реално време.

Машинно обучение

Машинно обучениеизследва как компютрите могат да учат (или да подобрят своята производителност) въз основа на данни. Основна изследователска област е компютърните програми да се научат автоматично да разпознават сложни модели и да вземат интелигентни решения въз основа на данни. Например, типичен проблем с машинното обучение е да се програмира компютър, така че да може автоматично да разпознава ръкописни пощенски кодове в пощата, след като се научи от набор от примери.

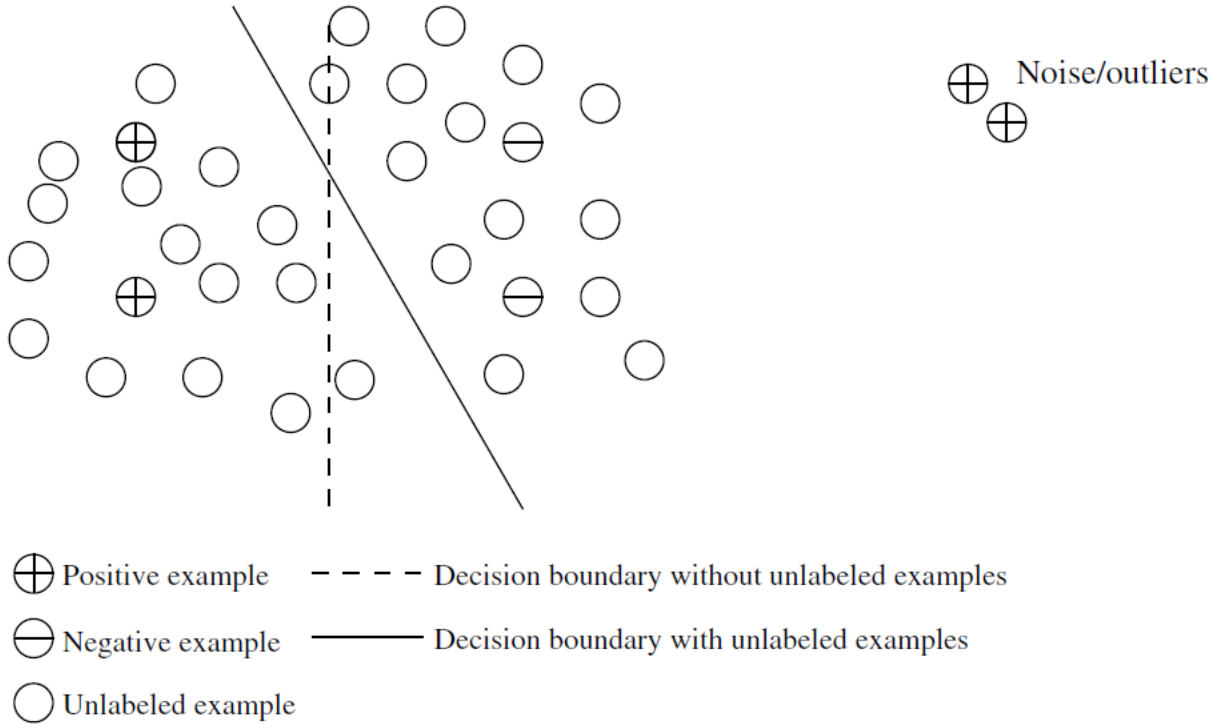
Машинното обучение е бързо развиваща се дисциплина. Тук илюстрираме класически проблеми в машинното обучение, които са тясно свързани с извличането на данни.

Учене под наблюдениее основно синоним на класификация. Надзорът в обучението идва от обозначените примери в набора от данни за обучение. Например, в проблема с разпознаването на пощенски код, набор от ръкописни изображения на пощенски код и съответните им машинночетими преводи се използват като примери за обучение, които контролират изучаването на класификационния модел.

Учене без надзоре по същество синоним на групиране. Процесът на обучение е без надзор, тъй като входните примери не са обозначени с клас. Обикновено можем да използваме групиране, за да открием класове в данните. Например метод за обучение без надзор може да приеме като вход набор от изображения на ръкописни цифри. Да предположим, че намира 10 групи от данни. Тези клъстери могат да съответстват съответно на 10-те различни цифри от 0 до 9. Въпреки това, тъй като данните за обучение не са етикетирани, наученият модел не може да ни каже семантичното значение на намерените клъстери.

Полуконтролирано обучениее клас техники за машинно обучение, които използват както маркирани, така и немаркирани примери, когато изучават модел. При един подход обозначените примери се използват за изучаване на модели на класове, а немаркираните примери се използват за прецизиране на границите между класовете. За проблем с два класа можем да мислим за набора от примери, принадлежащи към един клас, като положителни примери, а тези, принадлежащи към другия клас, като отрицателни примери.

На фигура 1.12, ако не вземем предвид немаркираните примери, пунктираната линия е границата на решение, която най-добре разделя положителните примери от отрицателните примери. Използвайки немаркираните примери, можем да прецизираме границата на решението до плътната линия. Освен това можем да открием, че двата положителни примера в горния десен ъгъл, макар и обозначени, вероятно са шум или отклонения.



Фиг. 1.12. Полуконтролирано обучение.

Активно обучениее подход за машинно обучение, който позволява на потребителите да играят активна роля в процеса на обучение. Подходът за активно обучение може да помоли потребител (напр. експерт по домейн) да етикетира пример, който може да бъде от набор от немаркирани примери или синтезиран от програмата за обучение. Целта е да се оптимизира качеството на модела чрез активно придобиване на знания от човешки потребители, като се има предвид ограничение за това колко примера могат да бъдат помолени да етикетират.

Можете да видите, че има много прилики между извличането на данни и машинното обучение. За задачи за класифициране и клъстериране изследванията на машинното обучение често се фокусират върху точността на модела. В допълнение към точността, изследванията за извличане на данни поставят силен акцент върху ефективността и мащабируемостта на методите за извличане на големи масиви от данни, както и върху начините за обработка на сложни типове данни и изследване на нови, алтернативни методи.

Системи за бази данни и DataWarehouses

Изследването на системите за бази данни се фокусира върху създаването, поддръжката и използването на бази данни за организации и крайни потребители. По-специално, изследователите на системи за бази данни са установили високо признати принципи в моделите на данни, езиците за заявки, методите за обработка и оптимизация на заявки, съхранение на данни и методи за индексиране и достъп. Системите за бази данни често са добре известни с високата си мащабируемост при обработката на много големи, относително структурирани набори от данни.

Много задачи за извличане на данни трябва да обработват големи набори от данни или дори данни в реално време, бърз поток. Следователно извличането на данни може да използва добре технологиите за мащабируеми бази данни за постигане на висока ефективност и мащабируемост на големи набори от данни. Освен това задачите за извличане на данни могат да се използват за разширяване на възможностите на съществуващите системи за бази данни, за да задоволят сложните изисквания за анализ на данни на напреднали потребители.

Последните системи за бази данни са изградили възможности за систематичен анализ на данни върху данни от база данни, използвайки съоръжения за съхранение на данни и извличане на данни. Складът за данни интегрира данни, произхождащи от множество източници и различни времеви рамки. Той консолидира данни в многомерно пространство, за да формира частично материализирани кубове с данни. Моделът на куба на данните не само улеснява OLAP в многомерни бази данни, но също така насърчава многомерното извличане на данни (вижте раздел 1.3.2).

Извличане на информация

Извличането на информация (IR) е наука за търсене на документи или информация в документи. Документите могат да бъдат текстови или мултимедийни и могат да се намират в мрежата. Разликите между традиционните системи за извличане на информация и бази данни са двойни: Извличането на информация предполага, че (1) търсените данни са неструктурирани; и (2) заявките се формират основно от ключови думи, които нямат сложни структури (за разлика от SQL заявките в системите с бази данни).

Типичните подходи за извличане на информация приемат вероятностни модели. Например, текстов документ може да се разглежда като торба с думи, тоест набор от думи, които се появяват в документа. Езиковият модел на документа е функцията за плътност на вероятността, която генерира пакета от думи в документа. Приликата между два документа може да се измери чрез приликата между съответните им езикови модели.

Освен това тема в набор от текстови документи може да бъде моделирана като вероятностно разпределение върху речника, което се нарича тематичен модел. Текстов документ, който може да включва една или няколко теми, може да се разглежда като смесица от множество тематични модели.

Чрез интегриране на модели за извличане на информация и техники за извличане на данни можем да намерим основните теми в колекция от документи и, за всеки документ в колекцията, основните включени теми.

Все по-големи количества текстови и мултимедийни данни са натрупани и предоставени онлайн поради бързия растеж на мрежата и приложения като цифрови библиотеки, цифрови правителства и информационни системи за здравеопазване. Тяхното ефективно търсене и анализ повдигнаха много предизвикателни проблеми в извличането на данни. Следователно извличането на текст и извличането на мултимедийни данни, интегрирани с методите за извличане на информация, стават все по-важни.

Кои видове приложения са насочени?

Където има данни, има и приложения за извличане на данни

Като дисциплина, ориентирана към приложенията, извличането на данни е отбелязало големи успехи в много приложения. Невъзможно е да се изброят

всички приложения, при които извличането на данни играе критична роля. Представянето на извличане на данни в области на приложение с интензивно знание, като биоинформатика и софтуерно инженерство, изисква по-задълбочено третиране и е извън обхвата на тази книга. За да демонстрираме значението на приложенията като основно измерение в изследванията и развитието на извличането на данни, ние накратко обсъждаме два изключително успешни и популярни примера за извличане на данни за извличане на данни: бизнес разузнаване и търсачки.

Бизнес разузнаване

За предприятията е от решаващо значение да придобият по-добро разбиране на търговския контекст на своята организация, като например своите клиенти, пазара, предлагането и ресурсите и конкурентите. Технологиите за бизнес разузнаване (BI) предоставят исторически, текущи и прогнозни изгледи на бизнес операциите. Примерите включват отчитане, онлайн аналитична обработка, управление на бизнес ефективността, конкурентно разузнаване, бенчмаркинг и прогнозни анализи.

„Колко важно е бизнес разузнаването?“ Без извличане на данни много фирми може да не са в състояние да извършват ефективен пазарен анализ, да сравняват отзивите на клиентите за подобни продукти, да откриват силните и слабите страни на своите конкуренти, да задържат много ценни клиенти и да вземат интелигентни бизнес решения.

Ясно е, че извличането на данни е в основата на бизнес разузнаването. Онлайн инструментите за аналитична обработка в бизнес разузнаването разчитат на съхранение на данни и многомерно извличане на данни. Техниките за класификация и прогнозиране са в основата на предсказуемия анализ в бизнес разузнаването, за който има много приложения при анализиране на пазари, доставки и продажби. Освен това клъстерирането играе централна роля в управлението на взаимоотношенията с клиентите, което групира клиентите въз основа на техните прилики. Използвайки техники за извличане на характеристики, можем да разберем по-добре характеристиките на всяка клиентска група и да разработим персонализирани програми за възнаграждение на клиентите.

Уеб търсачки

Уеб търсачката е специализиран компютърен сървър, който търси информация в мрежата. Резултатите от търсене на потребителска заявка често се връщат като списък (понякога наричан хитове). Попаденията могат да се състоят от уеб страници, изображения и други видове файлове. Някои търсачки също търсят и връщат данни, налични в публични бази данни или отворени директории.

Търсачките се различават от уеб директориите по това, че уеб директориите се поддържат от човешки редактори, докато търсачките работят алгоритмично или чрез комбинация от алгоритмично и човешко въвеждане.

Уеб търсачките са по същество много големи приложения за извличане на данни. Различни техники за извличане на данни се използват във всички аспекти на търсачките, вариращи от обхождане 5 (напр. решаване кои страници да бъдат обхождани и честотите на обхождане), индексиране (напр. избиране на страници за индексиране и решаване до каква степен да бъде индексиран конструиран) и търсене (напр. решаване как да бъдат класирани страниците, кои реклами да бъдат добавени и как резултатите от търсенето могат да бъдат персонализирани или направени „съобразени с контекста“).

Търсачките поставят големи предизвикателства пред извличането на данни. Първо, те трябва да обработват огромно и непрекъснато нарастващо количество данни. Обикновено такива данни не могат да бъдат обработени с помощта на една или няколко машини. Вместо това търсачките често трябва да използват компютърни облаци, които се състоят от хиляди или дори стотици хиляди компютри, които съвместно копаят огромното количество данни. Увеличаването на методите за извличане на данни в компютърни облаци и големи разпределени набори от данни е област за по-нататъшно изследване.

Второ, уеб търсачките често трябва да се справят с онлайн данни. Една търсачка може да си позволи да конструира модел офлайн върху огромни набори от данни. За да направи това, той може да създаде класификатор на заявка, който присвоява заявка за търсене към предварително дефинирани категории въз основа на темата на заявката (т.е. дали заявката за търсене „ябълка“ е предназначена да извлича информация за плод или марка компютри).

Независимо дали моделът е конструиран офлайн, приложението на модела онлайн трябва да бъде достатъчно бързо, за да отговаря на потребителски запитвания в реално време.

Друго предизвикателство е поддържането и постепенното актуализиране на модел на бързо нарастващи потоци от данни. Например може да се наложи класификаторът на заявки да се поддържа постепенно, тъй като непрекъснато се появяват нови заявки и предварително дефинирани категории и разпределението на данните може да се промени. Повечето от съществуващите методи за обучение на модели са офлайн и статични и следователно не могат да се използват в такъв сценарий.

Трето, уеб търсачките често трябва да се справят със заявки, които се задават само много малък брой пъти. Да предположим, че една търсачка иска да предостави контекстно ориентирани препоръки за заявки. Това означава, че когато потребител постави заявка, търсачката се опитва да разбере контекста на заявката, използвайки потребителския профил и неговата история на заявките, за да върне по-персонализирани отговори в рамките на малка част от секундата. Въпреки това, въпреки че общият брой на зададените запитвания може да бъде огромен, повечето от запитванията могат да бъдат зададени само веднъж или няколко пъти. Такива силно изкривени данни са предизвикателство за много методи за извличане на данни и машинно обучение.

Основни проблеми при извличането на данни

Животът е кратък, но изкуството е дълго. – Хипократ

Извличането на данни е динамична и бързо развиваща се област с големи силни страни. В този раздел накратко очертаваме основните проблеми в изследванията за извличане на данни, като ги разделяме на пет групи: методология за извличане на данни, взаимодействие с потребителите, ефективност и мащабируемост, разнообразие от типове данни и извличане на данни и общество. Много от тези въпроси са били разгледани в скорошни изследвания и разработки за извличане на данни до известна степен и сега се считат за изисквания за извличане на данни; други все още са на етап проучване. Проблемите продължават да стимулират по-нататъшно проучване и подобряване на извличането на данни.

Методология на копаене

Изследователите усилено разработват нови методологии за извличане на данни. Това включва изследване на нови видове знания, копаене в многомерно пространство, интегриране на методи от други дисциплини и разглеждане на семантичните връзки между обектите с данни. В допълнение, методологиите за копаене трябва да вземат предвид проблеми като несигурност на данните, шум и непълнота. Някои методи за копаене изследват как могат да се използват определени от потребителя мерки за оценка на интереса на откритите модели, както и за насочване на процеса на откриване. Нека да разгледаме тези различни аспекти на методологията за копаене.

*Копаене на различни и нови видове знания:*Извличането на данни обхваща широк спектър от задачи за анализ на данни и откриване на знания, от характеризиране на данни и дискриминация до анализ на асоциации и корелация, класификация, регресия, групиране, анализ на извънредни стойности, анализ на последователности и анализ на тенденциите и еволюцията. Тези задачи могат да използват една и съща база данни по различни начини и изискват разработването на множество техники за извличане на данни. Поради разнообразието от приложения продължават да се появяват нови задачи за копаене, което прави извличането на данни динамична и бързо развиваща се област. Например, за ефективно откриване на знания в информационни мрежи, интегрираното клъстериране и класиране може да доведе до откриването на висококачествени клъстери и обектни рангове в големи мрежи.

*Знания за копаене в многомерно пространство:*Когато търсим знания в големи набори от данни, можем да изследваме данните в многомерно пространство. Това означава, че можем да търсим интересни модели сред комбинации от измерения (атрибути) на различни нива на абстракция. Такова копаене е известно като (проучвателно) многоизмерно извличане на данни. В много случаи данните могат да бъдат агрегирани или разглеждани като многоизмерен куб с данни. Знанието за копаене в пространството на куба може значително да подобри мощността и гъвкавостта на извличането на данни.

*Извличане на данни - интердисциплинарно усилие:*Силата на извличането на данни може да бъде значително подобрена чрез интегриране на нови методи

от множество дисциплини. Например, за да извличаме данни с текст на естествен език, има смисъл да слеем методите за извличане на данни с методи за извличане на информация и обработка на естествен език. Като друг пример, помислете за копаене на софтуерни грешки в големи програми. Тази форма на копаене, известна като копаене на грешки, се възползва от включването на знания за софтуерно инженерство в процеса на копаене на данни.

*Увеличаване на силата на откриване в мрежова среда:*Повечето обекти с данни се намират в свързана или взаимосвързана среда, независимо дали става дума за уеб, връзки с бази данни, файлове или документи. Семантичните връзки между множество обекти с данни могат да се използват с предимство при извличането на данни. Знанието, получено в един набор от обекти, може да се използва за стимулиране на откриването на знания в „свързан“ или семантично свързан набор от обекти.

*Работа с несигурност, шум или непълнота на данните:*Данните често съдържат шум, грешки, изключения или несигурност или са непълни. Грешките и шумът могат да объркат процеса на извличане на данни, което води до извличане на грешни модели. Почистване на данни, предварителна обработка на данни, откриване и премахване на извънредни стойности и аргументиране на несигурността са примери за техники, които трябва да бъдат интегрирани в процеса на извличане на данни.

*Оценка на шаблони и копаене, ръководено от шаблони или ограничения:*Не всички модели, генерирани от процеси за извличане на данни, са интересни. Това, което прави модела интересен, може да варира от потребител до потребител. Следователно са необходими техники за оценка на интереса на откритите модели въз основа на субективни мерки. Те оценяват стойността на моделите по отношение на даден потребителски клас въз основа на потребителските вярвания или очаквания. Освен това, като използваме мерки за интересност или зададени от потребителя ограничения за насочване на процеса на откриване, можем да генерираме по-интересни модели и да намалим пространството за търсене.

Взаимодействие с потребителя

Потребителят играе важна роля в процеса на извличане на данни. Интересни области на изследване включват как да се взаимодейства със система за извличане на данни, как да се включат базовите познания на потребителя в добива и как да се визуализират и разбират резултатите от извличането на данни. Представяме всеки от тях тук.

Интерактивно копаене: Процесът на извличане на данни трябва да бъде силно интерактивен. Поради това е важно да се изградят гъвкави потребителски интерфейси и проучвателна среда за копаене, улесняваща взаимодействието на потребителя със системата. Потребителят може да пожелае първо да вземе извадка от набор от данни, да проучи общите характеристики на данните и да оцени потенциалните резултати от копаене. Интерактивното копаене трябва да позволява на потребителите динамично да променят фокуса на търсене, да прецизират заявките за копаене въз основа на върнатите резултати и да пробиват, заравят и въртят през пространството на данни и знания интерактивно, динамично изследвайки „кубното пространство“, докато копаят.

Включване на основни знания: Основните знания, ограниченията, правилата и друга информация относно изследваната област трябва да бъдат включени в процеса на откриване на знания. Такива знания могат да се използват за оценка на модели, както и за насочване на търсенето към интересни модели.

Ad hoc извличане на данни и езици за заявки за извличане на данни: Езиците за заявки (напр. SQL) са изиграли важна роля в гъвкавото търсене, защото позволяват на потребителите да задават ad hoc заявки. По подобен начин езиците за заявки за извличане на данни от високо ниво или други гъвкави потребителски интерфейси от високо ниво ще дадат на потребителите свободата да дефинират ad hoc задачи за извличане на данни. Това трябва да улесни спецификацията на съответните набори от данни за анализ, знанията за домейна, видовете знания, които трябва да бъдат извлечени, и условията и ограниченията, които да бъдат наложени върху откритите модели. Оптимизирането на обработката на такива гъвкави заявки за копаене е друга обещаваща област на изследване.

Представяне и визуализация на резултатите от извличане на данни: Как една система за извличане на данни може да представи резултатите от извличането на данни, ярко и гъвкаво, така че откритите знания да могат лесно да бъдат разбрани и директно използвани от хората? Това е особено важно, ако

процесът на извличане на данни е интерактивен. Тя изисква системата да възприеме експресивни представяния на знания, удобни за потребителя интерфейси и техники за визуализация.

Ефективност и мащабируемост

Ефективността и мащабируемостта винаги се вземат предвид при сравняване на алгоритми за извличане на данни. Тъй като количествата данни продължават да се умножават, тези два фактора са особено критични.

*Ефективност и мащабируемост на алгоритмите за извличане на данни:*Алгоритмите за извличане на данни трябва да бъдат ефективни и мащабируеми, за да извличат ефективно информация от огромни количества данни в много хранилища на данни или в динамични потоци от данни. С други думи, времето за изпълнение на алгоритъм за извличане на данни трябва да бъде предвидимо, кратко и приемливо от приложенията. Ефективност, мащабируемост, производителност, оптимизация и възможност за изпълнение в реално време са ключови критерии, които стимулират разработването на много нови алгоритми за извличане на данни.

*Алгоритми за паралелен, разпределен и инкрементален добив:*Огромният размер на много набори от данни, широкото разпространение на данни и изчислителната сложност на някои методи за извличане на данни са фактори, които мотивират разработването на паралелни и разпределени алгоритми за извличане на данни с интензивно използване на данни. Такива алгоритми първо разделят данните на „парчета“. Всяко парче се обработва паралелно чрез търсене на модели. Паралелните процеси могат да взаимодействат един с друг. Моделите от всеки дял в крайна сметка се обединяват.

*Облачни изчисления*и клъстерните изчисления, които използват компютри по разпределен и съвместен начин за справяне с много мащабни изчислителни задачи, също са активни изследователски теми в паралелното извличане на данни. В допълнение, високата цена на някои процеси за извличане на данни и нарастващият характер на входа насърчават поетапното извличане на данни, което включва нови актуализации на данни, без да се налага да копаете всички данни „от нулата“. Такива методи извършват постепенно модифициране на знанието, за да променят и засилят това, което е било открито преди това.

Разнообразие от типове бази данни

Голямото разнообразие от типове бази данни води до предизвикателства пред извличането на данни. Те включват

*Работа със сложни типове данни:*Разнообразни приложения генерират широк спектър от нови типове данни, от структурирани данни като релационни данни и данни от хранилище на данни до полуструктурирани и неструктурирани данни; от стабилни хранилища на данни към динамични потоци от данни; от обикновени обекти с данни до времеви данни, биологични последователности, сензорни данни, пространствени данни, хипертекстови данни, мултимедийни данни, софтуерен програмен код, уеб данни и данни от социални мрежи. Нереалистично е да се очаква една система за извличане на данни да извлича всички видове данни, като се има предвид разнообразието от типове данни и различните цели на извличането на данни. Системи за извличане на данни, посветени на домейн или приложение, се изграждат за задълбочено извличане на специфични видове данни. Изграждането на ефективни и ефикасни инструменти за извличане на данни за различни приложения остава предизвикателна и активна област на изследване.

*Динамични, мрежови и глобални хранилища за данни за копаене:*Множество източници на данни са свързани чрез Интернет и различни видове мрежи, образувайки гигантски, разпределени и разнородни глобални информационни системи и мрежи. Откриването на знания от различни източници на структурирани, полуструктурирани или неструктурирани, но взаимосвързани данни с разнообразна семантика на данни поставя големи предизвикателства пред извличането на данни. Извличането на такива гигантски, взаимосвързани информационни мрежи може да помогне за разкриването на много повече модели и знания в хетерогенни набори от данни, отколкото могат да бъдат открити от малък набор от изолирани хранилища на данни. Уеб копаене, извличане на данни от множество източници и извличане на информационни мрежи се превърнаха в предизвикателни и бързо развиващи се полета за извличане на данни.

1.7.5 Извличане на данни и общество

Как извличането на данни влияе на обществото? Какви стъпки може да предприеме извличането на данни, за да се запази поверителността на хората? Използваме ли извличане на данни в ежедневието си, без дори да знаем, че го правим? Тези въпроси повдигат следните въпроси:

*Социални въздействия на извличането на данни:*Тъй като извличането на данни навлиза в ежедневието ни, е важно да се изследва въздействието на извличането на данни върху обществото. Как можем да използваме технологията за извличане на данни в полза на обществото? Как можем да се предпазим от злоупотребата му? Неправилното разкриване или използване на данни и потенциалното нарушаване на личната неприкосновеност и правата за защита на данните са области на безпокойство, които трябва да бъдат разгледани.

*Извличане на данни за запазване на поверителността:*Извличането на данни ще подпомогне научните открития, управлението на бизнеса, възстановяването на икономиката и защитата на сигурността (напр. откриването в реално време на нарушители и кибератаки). Това обаче крие риск от разкриване на лична информация на дадено лице. Проучванията за запазване на поверителността на публикуване на данни и извличане на информация продължават. Философията е да се наблюдава чувствителността на данните и да се запази поверителността на хората, докато се извършва успешно извличане на данни.

Невидимо извличане на данни: Не можем да очакваме всеки в обществото да научи и да овладее техники за извличане на данни. Все повече и повече системи трябва да имат вградени функции за извличане на данни, така че хората да могат да извършват извличане на данни или да използват резултати от извличане на данни просто чрез щракване с мишката, без никакви познания за алгоритми за извличане на данни. Интелигентните търсачки и интернет базираните магазини извършват такова невидимо извличане на данни, като включват извличане на данни в своите компоненти, за да подобрят тяхната функционалност и производителност. Това често се прави без знанието на потребителя. Например, когато купуват артикули онлайн, потребителите може да не знаят, че магазинът вероятно събира данни за моделите на купуване на своите клиенти, които могат да бъдат използвани, за да препоръчват други артикули за покупка в бъдеще.

Тези въпроси и много други, свързани с изследването, разработването и прилагането на извличане на данни, се обсъждат в цялата книга.

Резюме

Необходимостта е майка на изобретението.С нарастващия растеж на данните във всяко приложение, извличането на данни отговаря на непосредствената нужда от ефективен, мащабируем и гъвкав анализ на данни в нашето общество. Извличането на данни може да се разглежда като естествена еволюция на информационните технологии и сливане на няколко свързани дисциплини и области на приложение.

*Извличането на данни е процес на откриване на интересни модели*от огромни количества данни. Като процес на откриване на знания, той обикновено включва почистване на данни, интегриране на данни, избор на данни, трансформация на данни, откриване на модели, оценка на модели и представяне на знания.

*Един модел е интересен, ако е валиден върху тестови данни*с известна степен на сигурност, нов, потенциално полезен (напр. може да се действа или потвърждава предположение, за което потребителят е любопитен) и лесно разбираем от хората. Интересни модели представляват знание. Мерките за интересност на модела, обективни или субективни, могат да се използват за насочване на процеса на откриване.

Представяме многоизмерен изглед на извличането на данни. Основните измерения са данни, знания, технологии и приложения.

*Извличането на данни може да се извършива върху всякакъв вид данни*стига данните да са значими за целево приложение, като данни от бази данни, данни от хранилище на данни, транзакционни данни и разширени типове данни. Разширените типове данни включват свързани с времето или последователни данни, потоци от данни, пространствени и пространствено-времеви данни, текстови и мултимедийни данни, графични и мрежови данни и уеб данни.

*Складът за данни е хранилище за дългосрочно съхранение на данни*от множество източници, организирани така, че да улесняват вземането на управленски решения. Данните се съхраняват под унифицирана схема и обикновено са обобщени. Системите за съхранение на данни предоставят

възможности за многоизмерен анализ на данни, наричани общо онлайн аналитична обработка.

Многомерно извличане на данни (наричан още проучвателно многоизмерно извличане на данни) интегрира основни техники за извличане на данни с базиран на OLAP многоизмерен анализ. Той търси интересни модели сред множество комбинации от измерения (атрибути) на различни нива на абстракция, като по този начин изследва многоизмерното пространство на данните.

Функции за извличане на данни се използват за определяне на видовете модели или знания, които да бъдат намерени в задачите за извличане на данни. Функционалностите включват характеризиране и дискриминация; извличането на чести модели, асоциации и корелации; класификация и регресия; клъстерен анализ; и откриване на отклонения. Тъй като продължават да се появяват нови типове данни, нови приложения и нови изисквания за анализ, няма съмнение, че ще виждаме все повече и повече нови задачи за извличане на данни в бъдеще.

Извличане на данни, като силно управляван от приложения домейн, включва технологии от много други области. Те включват статистика, машинно обучение, бази данни и системи за съхранение на данни и извличане на информация. Интердисциплинарният характер на изследването и развитието на извличането на данни допринася значително за успеха на извличането на данни и неговите обширни приложения.

*Извличане на данни*има много успешни приложения, като бизнес разузнаване, уеб търсене, биоинформатика, здравна информатика, финанси, цифрови библиотеки и цифрови правителства.

*Има много***предизвикателен**проблеми в изследванията за извличане на данни. Областите включват методология за копаене, взаимодействие с потребителите, ефективност и мащабируемост и работа с различни типове данни. Изследванията за извличане на данни оказаха силно въздействие върху обществото и ще продължат да го правят в бъдеще.

Упражнения

1.1Какво е извличане на данни? В отговора си обърнете внимание на следното:

(а) Друга реклама ли е?

Проект BG05M2OP001-2.016-0003 „Модернизация на Национален военен университет "Васил Левски"- гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, финансиран от ниски съюз чрез Европейските структурни и инвестиционни фондове.

(б) Това проста трансформация или приложение на технология, разработена от бази данни, статистика, машинно обучение и разпознаване на образи?

(c) Представихме мнение, че извличането на данни е резултат от еволюцията на технологията за бази данни. Смятате ли, че извличането на данни също е резултат от еволюцията на изследванията на машинното обучение? Можете ли да представите такива възгледи въз основа на историческия прогрес на тази дисциплина? Адресирайте същото за полетата статистика и разпознаване на образи.

(d) Опишете стъпките, включени в извличането на данни, когато се разглежда като процес на откриване на знания.

1.2Как се различава хранилището на данни от базата данни? По какво си приличат?

1.3Дефинирайте всяка от следните функционалности за извличане на данни: характеризиране, дискриминация, асоциационен и корелационен анализ, класификация, регресия, групиране и анализ на отклонения. Дайте примери за всяка функционалност за извличане на данни, като използвате база данни от реалния живот, с която сте запознати.

1.4Представете пример, при който извличането на данни е от решаващо значение за успеха на даден бизнес. От какви функции за извличане на данни се нуждае този бизнес (напр. помислете за видовете модели, които могат да бъдат извличани)? Могат ли такива модели да бъдат генерирани алтернативно чрез обработка на заявки за данни или прост статистически анализ?

1.5Обяснете разликата и приликата между дискриминация и класификация, между характеризиране и групиране и между класификация и регресия.

1.6Въз основа на вашите наблюдения опишете друг възможен вид знание, което трябва да бъде открито чрез методи за извличане на информация, но не е изброено в тази глава. Изисква ли методология за копаене, която е доста различна от описаните в тази глава?

1.7*Извънредни стойности*често се отхвърлят като шум. Въпреки това, боклукът на един човек може да бъде съкровище за друг. Например изключенията при транзакции с кредитни карти могат да ни помогнат да открием

измамното използване на кредитни карти. Използвайки откриването на измама като пример, предложете два метода, които могат да се използват за откриване на отклонения и обсъдете кой е по-надежден.

1.8Опишете три предизвикателства пред извличането на данни по отношение на методологията за извличане на данни и проблеми с взаимодействието с потребителите.

1.9Какви са основните предизвикателства при извличането на огромно количество данни (напр. милиарди кортежи) в сравнение с извличането на малко количество данни (напр. набор от данни от няколкостотин кортежа)?

1.10Очертайте основните изследователски предизвикателства на извличането на данни в една конкретна област на приложение, като анализ на данни от поток/сензор, анализ на пространствено-времеви данни или биоинформатика.