МЕТОДОЛОГИЯ ЗА ИЗВЛИЧАНЕ НА ЗНАНИЯ ОТ ДАННИ

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.

Извличане 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 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 ОТ 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.

Извличане на знания от данни и обществените реакции

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 извършва 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 **предизвикателни 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 вид реклама ли е?

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

(с) 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.

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

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

1.3 Дефинирайте 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.

1.4 Представете 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?

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

1.6 Въз 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.

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

1.9 Какви 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)?

1.10 Очертайте 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.