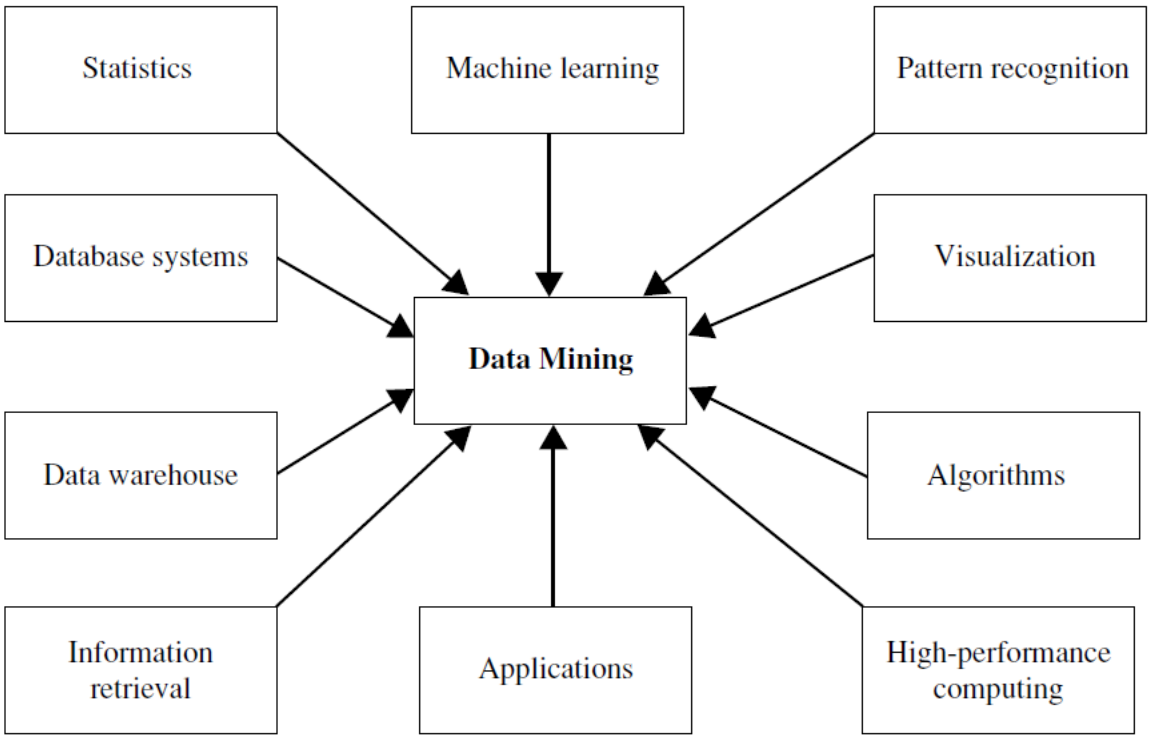
ТЕХНОЛОГИИ ЗА ЗИВЛИЧАНЕ НА ЗНАНИЯ ОТ ДАННИ

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.

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 обучение изследва 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 наблюдение е 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 машиночетими преводи се използват като примери за обучение, които контролират изучаването на класификационния модел.

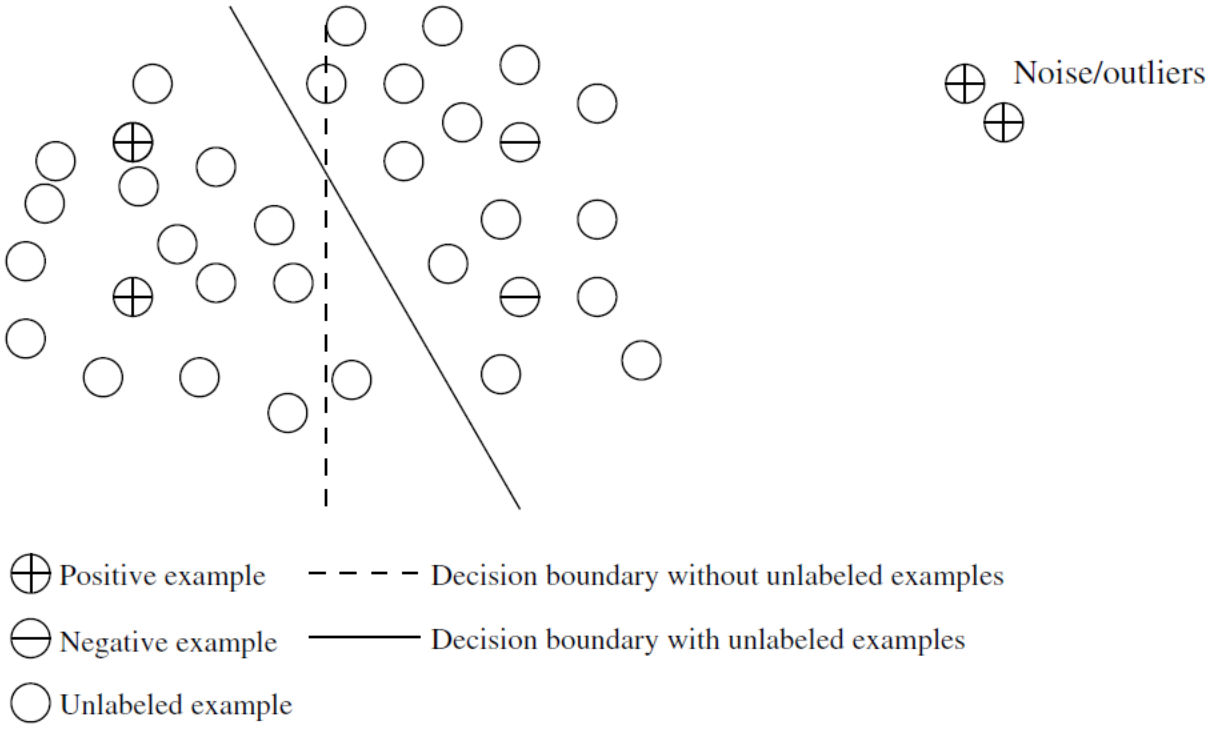
Учене без надзоре по същество синоним на групиране. Процесът на обучение е без надзор, тъй като входните примери не са обозначени с клас. Обикновено можем да използваме групиране, за да открием класове в данните.

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 обучение е 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.

Активно обучение е 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.



Фиг. 1.12. Полуконтролирано обучение.

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 управление на бази данни и хранилища за данни (Data Warehouses)

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 търсещи машини

Уеб търсещата машина 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 вечно“ – 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.