Data mining: torturant les dades fins que confessin[*]

Luis Carlos Molina Félix

Coordinador del programa de *Data mining* (UOC)
lmolinaf@uoc.edu

Resum: El títol d'aquest article és una explicació informal de l'activitat que realitza una tecnologia anomenada *data mining* (mineria de dades). El que es pretén amb aquesta tecnologia és descobrir coneixement *ocult* a partir de grans volums de dades. Des de la dècada passada, a causa dels grans avenços computacionals, s'ha anat incorporant a les organitzacions i ha acabat esdevenint un suport essencial per al moment de prendre decisions. Organitzacions com ara empreses, clubs professionals esportius, universitats i governs, entre altres, fan ús d'aquesta tecnologia com a *ajuda* en la presa de decisions. Alguns d'aquests exemples seran esmentats en el present treball.

1. Introducció

Cada dia generem una gran quantitat d'informació, algunes vegades conscients del fet que ho fem i, altres, inconscients, perquè ho desconeixem. Ens adonem que generem informació quan registrem la nostra entrada en el treball, quan entrem en un servidor per veure el nostre correu, quan paguem amb una targeta de crèdit o quan reservem un bitllet d'avió. Altres vegades no ens adonem que generem informació, com passa quan conduïm per una via on es comptabilitza el nombre d'automòbils que passen per minut, quan se segueix la nostra navegació per Internet o quan ens fan una fotografia del rostre en passar prop d'una oficina governamental.

Amb quina finalitat volem generar informació? Són molts els motius que ens porten a generar informació, ja que ens poden ajudar a controlar, optimitzar, administrar, examinar, investigar, planificar, predir, sotmetre, negociar o prendre decisions de qualsevol àmbit segons el domini en què ens movem. La informació en si mateixa és considerada un bé patrimonial. D'aquesta manera, el fet que una empresa tingui una pèrdua total o parcial d'informació provoca bastants perjudicis. És evident que la informació ha de ser protegida, però també explotada.

Què ens ha permès poder generar tanta informació? En els darrers anys, a causa del desenvolupament tecnològic a nivells exponencials tant en l'àrea de càlcul com en la de transmissió de dades, ha estat possible que es gestioni millor el maneig i l'emmagatzematge de la informació. Sens dubte, hi ha quatre factors importants que ens han portat a aquesta situació:

1. L'abaratiment dels sistemes d'emmagatzematge tant temporal com permanent.

2. L'increment de les velocitats de càlcul en els processadors.

*** Les transparències d'aquest article es poden obtenir a: http://www.lsi.upc.es/~lcmolina/about.htm[url1].**

Títol original: *Data mining: torturando a los datos hasta que confiesen*

3. Les millores en la confiabilitat i l'augment de la velocitat en la transmissió de dades.

4. El desenvolupament de sistemes administradors de bases de dades més potents.

Actualment tots aquests avantatges ens han menat a abusar de l'emmagatzematge de la informació en les bases de dades. Podem dir que algunes empreses emmagatzemen un cert tipus de dades que hem anomenat *dada-escriptura*, ja que només es desa (o *escriu*) al disc dur, però mai no se'n fa ús. Generalment, totes les empreses fan servir una dada anomenada *dada-escriptura-lectura*, que utilitzen per a fer consultes dirigides. Un nou tipus de dada a la qual hem anomenat *dada-escriptura-lectura-anàlisi* és la que proporciona en conjunt un vertader coneixement i ens dóna suport en la presa de decisions. Cal disposar de tecnologies que ens ajudin a explotar el potencial d'aquest tipus de dades.

La quantitat d'informació que ens arriba cada dia és tan immensa que ens resulta difícil assimilar-la. N'hi ha prou d'anar al cercador Altavista[url2] i sol·licitar la paraula *information* per veure que hi ha 171.769.416 llocs on ens poden dir alguna cosa sobre això. Suposant que disposem d'un minut per a veure el contingut de cada pàgina, trigaríem 326 anys a visitar-les totes. Això és impossible i, per tant, hi ha una clara necessitat de disposar de tecnologies que ens ajudin en els nostres processos de cerca i, encara més, de tecnologies que ens ajudin a comprendre'n el contingut.

El *data mining* sorgeix com una tecnologia que intenta ajudar a entendre el contingut d'una base de dades. En general, les *dades* són la matèria primera bruta. En el moment en què l'usuari els atribueix algun significat especial esdevenen *informació*. Quan els especialistes elaboren un model o en troben un, i fan que la interpretació de l'acarament entre la informació i aquell model representi un valor agregat, llavors ens referim a *coneixement*. A la figura 1 s'il·lustra la jerarquia que hi ha en una base de dades entre dada, informació i coneixement (Molina, 1998). S'hi observa igualment el volum que presenta en cada nivell i el valor que els responsables de les decisions hi donen en aquesta jerarquia. L'àrea interna dins el triangle representa els objectius que s'han proposat. La separació del triangle representa la unió estreta entre dada i informació, però no entre la informació i el coneixement. El *data mining* treballa al nivell superior buscant patrons, comportaments, agrupacions, seqüències, tendències o associacions que puguin generar algun model que ens permeti comprendre millor el domini per a *ajudar* en una possible presa de decisió.

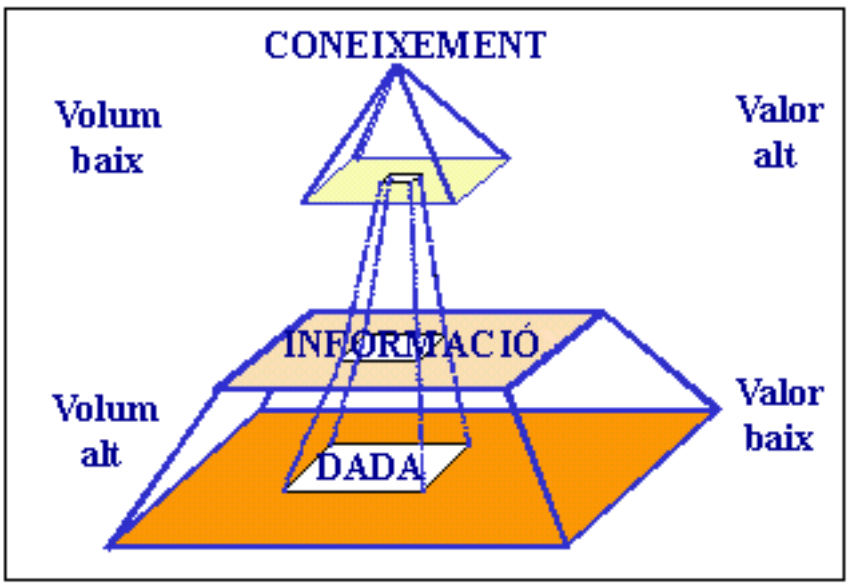

Figura 1. Relació entre dada, informació i coneixement (Molina, 1998).

2. *Data mining*: conceptes i història

Encara que des d'un punt de vista acadèmic el terme *data mining* és una etapa dins un procés més gran anomenat *extracció de coneixement en bases de dades* (*Knowledge Discovery in Databases* o KDD) en l'àmbit comercial, i també en aquest treball, ambdós termes s'utilitzen de manera indistinta. El que de veritat fa el *data mining* és aplegar els avantatges de diverses àrees com ara l'Estadística, la Intel·ligència Artificial, la Computació Gràfica, les Bases de Dades i el Processament Massiu, sobretot fent servir les bases de dades com a matèria primera. Una definició tradicional és la següent: "Un procés no trivial d'identificació vàlida, nova, potencialment útil i entenedora de patrons comprensibles que es troben ocults en les dades" (Fayyad i altres, 1996). Des del nostre punt de vista, el definim com "la integració d'un conjunt d'àrees que tenen com a propòsit la identificació d'un coneixement obtingut a partir de les bases de dades que aportin un biaix cap a la presa de decisió" (Molina i altres, 2001).

La idea de *data mining* no és nova. Ja des dels anys seixanta els estadístics manejaven termes com ara *data fishing*, *data mining* o *data archaeology* amb la idea de trobar correlacions sense una hipòtesi prèvia en bases de dades amb soroll. A començament dels anys vuitanta, Rakesh Agrawal, Gio Wiederhold, Robert Blum, Gregory Piatetsky-Shapiro, entre altres, van començar a consolidar els termes de *data mining* i KDD.[3] A final dels anys vuitanta només existien un parell d'empreses dedicades a aquesta tecnologia; el 2002 hi ha més de 100 empreses al món que ofereixen al voltant de 300 solucions. Les llistes de discussió sobre aquest tema les formen investigadors de més de vuitanta països. Aquesta tecnologia ha estat un bon punt de trobada entre persones pertanyents a l'àmbit acadèmic i al dels negocis.

El *data mining* és una tecnologia formada per etapes que integra diverses àrees i que no s'ha de confondre amb un gran programari. Durant el desenvolupament d'un projecte d'aquest tipus es fan servir diferents aplicacions en cada etapa, que poden ser estadístiques, de visualització de dades o d'intel·ligència artificial, principalment. Actualment hi ha aplicacions o eines comercials de *data mining* molt poderoses que contenen una infinitat d'utilitats que faciliten el desenvolupament d'un projecte. Tanmateix, gairebé sempre s'acaben complementant amb una altra eina.

3. Aplicacions d'ús

Cada any en els diferents congressos, simposis i tallers que s'organitzen al món es reuneixen investigadors amb aplicacions molt diverses. Sobretot als Estats Units, el *data mining* s'ha anat incorporant a la vida d'empreses, governs, universitats, hospitals i diverses organitzacions que estan interessades a explorar les seves bases de dades.

Podem dir que "en *data mining* cada cas és un cas". Tanmateix, en termes generals, el procés es compon de quatre etapes principals:

1. Determinació dels objectius. Tracta de la delimitació dels objectius que el *client* desitja a partir de l'orientació de l'especialista en *data mining*.

2. Preprocessament de les dades. Fa referència a la selecció, la neteja, l'enriquiment, la reducció i la transformació de les bases de dades. Aquesta etapa consumeix generalment al voltant del setanta per cent del temps total d'un projecte de *data mining*.

3. Determinació del model. Comença amb unes anàlisis estadístiques de les dades, i després se'n fa una visualització gràfica per tal de tenir-ne una primera aproximació. Segons els objectius plantejats i la tasca que s'ha de portar a terme, es poden utilitzar algoritmes desenvolupats en diferents àrees de la Intel·ligència Artificial.

3. Més detalls a http://www.kdnuggets.com[url3].

Títol original: *Data mining: torturando a los datos hasta que confiesen*

4. Anàlisi dels resultats. Verifica si els resultats obtinguts són coherents i els acara amb els obtinguts per les anàlisis estadístiques i de visualització gràfica. El *client* determina si són nous i si li aporten un nou coneixement que li permeti considerar les seves decisions.

A continuació es descriuen diversos exemples en els quals s'ha aplicat el *data mining*. S'han seleccionat de diversos dominis i amb diversos objectius per a observar-ne el potencial. Respecte dels models intel·ligents, s'ha comprovat que s'hi fan servir principalment arbres i regles de decisió, regles d'associació, xarxes neuronals, xarxes bayesianes, conjunts aproximats (*rough sets*), algoritmes d'agrupació (*clustering*), màquines de suport vectorial, algoritmes genètics i lògica difusa.

3.1. Al govern

L'FBI analitzarà les bases de dades comercials per a detectar terroristes.

A començament del mes de juliol de 2002 el director del Federal Bureau of Investigation (FBI), John Aschcroft, va anunciar que el Departament de Justícia començarà a introduir-se en la vasta quantitat de dades comercials referents als hàbits i preferències de compra dels consumidors, a fi de descobrir terroristes potencials abans que executin una acció.[4] Alguns experts asseguren que, amb aquesta informació, l'FBI unirà totes les bases de dades probablement mitjançant el número de la Seguretat Social , cosa que permetrà saber si una persona fuma, quina talla i tipus de roba fa servir, el seu registre d'arrests, el seu salari, les revistes a què està subscrit, la seva alçada i el seu pes, les seves contribucions a l'Església, grups polítics o organitzacions no governamentals, les seves malalties cròniques (com ara diabetis o asma), els llibres que llegeix, els productes de supermercat que compra, si ha fet classes de vol o si té comptes de banc oberts, entre altres.[5] La inversió inicial ronda els setanta milions de dòlars estatunitencs per a consolidar els magatzems de dades, desenvolupar xarxes de seguretat amb vista a compartir informació i implementar nou programari analític i de visualització.

3.2. A l'empresa

Detectar fraus en les targetes de crèdit.

En l'any 2001 les institucions financeres de tot el món van perdre més de 2.000 milions de dòlars estatunitencs en fraus amb targetes de crèdit i dèbit. El Falcon Fraud Manager[6] és un sistema intel·ligent que examina transaccions, propietaris de targetes i dades financeres amb l'objectiu de detectar i mitigar fraus. Al principi estava pensat, en institucions financeres de l'Amèrica del Nord, per a detectar fraus en targetes de crèdit. Tanmateix, actualment se li han incorporat funcionalitats d'anàlisi en les targetes comercials, de combustibles i de dèbit.[7] El sistema Falcon ha permès estalviar més de sis-cents milions de dòlars estatunitencs cada any i protegeix aproximadament més de quatre-cents cinquanta milions de pagaments amb targeta a tot el món –aproximadament el seixanta-cinc per cent de totes les transaccions amb targeta de crèdit.

Descobrir el perquè de la deserció de clients d'una empresa operadora de telefonia mòbil.

Aquest estudi va ser desenvolupat en una operadora espanyola que bàsicament va situar els seus objectius en dos punts: l'anàlisi del perfil dels clients que es donen de baixa i la predicció del comportament dels seus nous clients. Es van analitzar els

4. Vegeu-ne més a http://www.fcw.com/fcw/articles/2002/0603/news-fbi-06-03-02.asp[url4].
5. Vegeu-ne més a http://tierra.ucsd.edu/archives/ats-l/2002.06/msg00013.html[url5].
6. Vegeu-ne més a http://www.fairisaac.com/page.cfm/press_id=325[url6].
7. American Express ha aconseguit entre un deu i un quinze per cent d'increment en l'ús de les seves targetes ajudant-se de tècniques de *data mining*.

Títol original: *Data mining: torturando a los datos hasta que confiesen*

diferents historials de clients que havien abandonat l'operadora (12,6%) i de clients que continuaven amb el seu servei (87,4%). També es van analitzar les variables personals de cada client (estat civil, edat, sexe, nacionalitat, etc.). De la mateixa manera es van estudiar, per a cada client, la morositat, la freqüència i l'horari d'ús del servei, els descomptes i el percentatge de trucades locals, interprovincials, internacionals i gratuïtes. Al contrari del que es podria pensar, els clients que abandonaven l'operadora generaven guanys per a l'empresa; ara bé, una de les conclusions més importants que se'n va extreure va ser que els clients que es donaven de baixa rebien poques promocions i registraven un nombre més alt d'incidències respecte de la mitjana. D'aquesta manera es va recomanar a l'operadora fer un estudi sobre les seves ofertes i analitzar profundament les incidències rebudes per aquests clients. En descobrir el perfil que presentaven, l'operadora va haver de dissenyar un tracte més personalitzat per als seus clients actuals amb aquestes característiques. Amb l'objectiu de prevenir el comportament dels nous clients es va dissenyar un sistema de predicció basat en la quantitat de dades que es podia obtenir dels nous clients en comparació amb el comportament de clients anteriors.

Preveure el volum de les audiències televisives.

La British Broadcasting Corporation (BBC) del Regne Unit empra un sistema destinat a predir el volum de les audiències televisives per a un programa proposat, i també el temps òptim d'exhibició (Brachman i altres, 1996). El sistema utilitza xarxes neuronals i arbres de decisió aplicats a dades històriques de la cadena per a determinar els criteris que hi participen segons el programa que s'ha de presentar.[8] La versió final funciona tan bé com un expert humà, amb l'avantatge que s'adapta més fàcilment als canvis perquè és reentrenada constantment amb dades actuals.

3.3. A la universitat

Saber si els nous llicenciats d'una universitat duen a terme activitats professionals relacionades amb els seus estudis.

Es va fer un estudi sobre els nous llicenciats de la carrera d'Enginyeria en Sistemes Computacionals de l'Institut Tecnològic de Chihuahua II,[9] a Mèxic (Rodas, 2001). Es volia observar si els nous titulats s'introduïen en activitats professionals relacionades amb els seus estudis i, en cas negatiu, es mirava de conèixer el perfil que caracteritzava els exalumnes durant la seva estada a la universitat. L'objectiu era saber si amb els plans d'estudi de la universitat i l'aprofitament de l'alumne es feia una bona inserció laboral o si existien altres variables que participaven en el procés. En la informació considerada hi havia el sexe, l'edat, l'escola de procedència, el rendiment acadèmic, la zona econòmica on es tenia l'habitatge i l'activitat professional, entre altres variables. Mitjançant l'aplicació de conjunts aproximats es va descobrir que hi havia quatre variables que determinaven l'adequada inserció laboral, les quals se citen segons la seva importància: zona econòmica on habitava l'estudiant, col·legi d'on provenia, nota d'entrada i mitjana final en sortir de la carrera. A partir d'aquests resultats, la universitat haurà de fer un estudi socioeconòmic sobre grups d'alumnes que pertanyen a les classes econòmiques baixes per tal de donar possibles solucions, ja que tres de les quatre variables no depenien de la universitat.

3.4. En investigacions espacials

Projecte SKYCAT.

Durant sis anys el Second Palomar Observatory Sky Survey (POSS-II) va reunir tres terabytes d'imatges que contenien aproximadament dos milions d'objectes al cel. Es van

8. Més detalls a http://www.mining.dk/SPSS/Nyheder/nr7case_bbc.htm[url7].
9. Equivalent, a Espanya, a una universitat politècnica.

digitalitzar tres mil fotografies a una resolució de 16 bits per píxel amb 23.040 x 23.040 píxels per imatge. L'objectiu era formar un catàleg de tots aquells objectes. El sistema Sky Image Cataloguing and Analysis Tool (SKYCAT) es basa en tècniques d'agrupació (*clustering*) i arbres de decisió per a poder classificar els objectes en estrelles, planetes, sistemes, galàxies, etc. amb una alta confiabilitat (Fayyad i altres, 1996). Els resultats han ajudat els astrònoms a descobrir setze nous quàsars amb corriment cap al vermell que els inclou entre els objectes més llunyans de l'univers i, per tant, més antics. Aquests quàsars són difícils de trobar i permeten saber més coses sobre els orígens de l'univers.

3.5. Als clubs esportius

L'AC de Milà utilitza un sistema intel·ligent per a prevenir lesions.

Aquesta temporada el club començarà a fer servir xarxes neuronals per a prevenir lesions i optimitzar el condicionament de cada atleta. Això ajudarà a seleccionar el fitxatge d'un possible jugador o a alertar el metge de l'equip d'una possible lesió.[10] El sistema, creat per Computer Associates International, és alimentat per dades de cada jugador, relacionades amb el seu rendiment, l'alimentació i la resposta a estímuls externs, que s'obtenen i s'analitzen cada quinze dies. El jugador duu a terme determinades activitats que són controlades per vint-i-quatre sensors connectats al cos i que transmeten senyals de ràdio que posteriorment són emmagatzemats en una base de dades. Actualment el sistema disposa de 5.000 casos registrats que permeten prevenir alguna possible lesió. Amb això, el club intenta estalviar diners evitant comprar jugadors que presentin una alta probabilitat de lesió, cosa que faria inevitable la renegociació del seu contracte. D'altra banda, el sistema pretén trobar les diferències entre les lesions d'atletes d'ambdós sexes, i també saber si una determinada lesió es relaciona amb l'estil de joc d'un país concret on es practica el futbol.

Els equips de l'NBA utilitzen aplicacions intel·ligents per a donar suport al cos d'entrenadors.

L'Advanced Scout[11] és un programari que empra tècniques de *data mining* i que han desenvolupat investigadors d'IBM per a detectar patrons estadístics i esdeveniments *rars*. Té una interfície gràfica molt amigable orientada a un objectiu molt específic: analitzar el joc dels equips de la National Basketball Association (NBA).

El programari utilitza tots els registres desats de cada esdeveniment en cada joc: passis, encistellades, rebots i marcatge doble (*double team*) a un jugador per part de l'equip contrari, entre altres. L'objectiu és ajudar els entrenadors a aïllar situacions que no detecten quan observen el joc en viu o en pel·lícula.

Un dels resultats més interessants va ser un que fins aleshores els entrenadors dels Knicks de Nova York no havien observat. El marcatge doble a un jugador pot donar generalment l'oportunitat a un altre jugador d'encistellar més fàcilment. Tanmateix, es va veure que, quan els Bulls de Chicago jugaven contra els Knicks, el percentatge d'encistellades després que al centre dels Knicks, Patrick Ewing, li fessin marcatge doble era extremament baix, cosa que indicava que els Knicks no reaccionaven correctament als marcatges dobles. Amb l'objectiu de conèixer-ne el perquè, el cos d'entrenadors va estudiar amb detall totes les pel·lícules de partits contra Chicago. Van observar que els jugadors de Chicago trencaven el seu marcatge doble molt ràpid de tal manera que podien tapar l'encistellador lliure dels Knicks abans de preparar-se per efectuar el seu tir. Amb aquest coneixement, els entrenadors van crear estratègies alternatives per a fer front al marcatge doble.

La temporada passada IBM va oferir l'Advanced Scout a l'NBA, que es va convertir així en un patrocinador corporatiu. L'NBA va donar als seus vint-i-nou equips l'oportunitat

10. Més detalls a http://www.msnbc.com/news/756968.asp[url8].
11. Vegeu http://domino.research.ibm.com/comm/wwwr_thinkresearch.nsf/pages/datamine296.html[url9].

Títol original: *Data mining: torturando a los datos hasta que confiesen*

d'aplicar-lo. Divuit equips ho fan en aquests moments, amb uns resultats interessants.

4. Extensions del *data mining*

4.1. *Web mining*

Una de les extensions del *data mining* consisteix a aplicar les seves tècniques a documents i serveis del web; és el que s'anomena *web mining* (mineria de web) (Kosala i altres, 2000). Tothom que visita un lloc a Internet hi deixa empremtes digitals (adreces IP, navegador, galetes, etc.) que els servidors automàticament emmagatzemen en una bitàcola d'accessos (*log*). Les eines de *web mining* analitzen aquests *logs* i els processen per a produir informació significativa, per exemple, com és la navegació d'un client abans de fer una compra en línia. Com que els continguts d'Internet estan formats per diversos tipus de dades, com ara text, imatge, vídeo, metadades o enllaços, investigacions recents fan servir el terme *multimedia data mining* (mineria de dades multimèdia) com una instància del *web mining* (Zaiane i altres, 1998) per a tractar aquest tipus de dades. Els accessos totals per domini, horaris d'accessos més freqüents i visites per dia, entre altres dades, són registrats per eines estadístiques que complementen tot el procés d'anàlisi del *web mining*.

Normalment, el *web mining* es pot classificar en tres dominis d'extracció de coneixement d'acord amb la naturalesa de les dades:

1. *Web content mining* (mineria de contingut web). És el procés que consisteix en l'extracció de coneixement del contingut de documents o les seves descripcions. La localització de patrons al text dels documents, el descobriment del recurs basat en conceptes d'indexació o la tecnologia basada en agents també poden formar part d'aquesta categoria.

2. *Web structure mining* (mineria d'estructura web). És el procés d'inferir coneixement de l'organització del WWW i l'estructura dels seus enllaços.

3. *Web usage mining* (mineria d'ús web). És el procés d'extracció de models interessants fent servir els *logs* dels accessos al web.

Alguns dels resultats que es poden obtenir després de l'aplicació dels diferents mètodes de *web mining* són:

- El vuitanta-cinc per cent dels clients que accedeixen a */productos/home.html* i a */productos/noticias.html* accedeixen també a */productos/historias_suceso.html*. Això pot indicar que hi ha alguna notícia interessant de l'empresa que fa que els clients vagin a històries d'esdeveniment. Igualment, aquest resultat podria permetre detectar la notícia destacada i col·locar-la potser a la pàgina principal de l'empresa.

- Els clients que fan una compra en línia cada setmana a */compra/producto1.html* tendeixen a ser de sectors del govern. Això podria menar a la proposta de diverses ofertes a aquest sector amb l'objectiu de potenciar més les seves compres.

- El seixanta per cent dels clients que van fer una compra en línia a */compra/producto1.html* també van comprar a */compra/producto4.html* al cap d'un mes. Això indica que es podria recomanar a la pàgina del producte 1 comprar el producte 4 i estalviar-se el cost de tramesa d'aquest producte.

Títol original: *Data mining: torturando a los datos hasta que confiesen*

Els exemples anteriors ens ajuden a fer-nos una petita idea del que podem obtenir. Tanmateix, a la realitat hi ha eines de mercat molt poderoses amb mètodes diversos i visualitzacions gràfiques excel·lents. Per a més informació, vegeu Mena (1999).

4.2. *Text mining*

Estudis recents indiquen que el vuitanta per cent de la informació d'una empresa està emmagatzemada en forma de documents. Sens dubte, aquest camp d'estudi és molt vast, per la qual cosa tècniques com la categorització de text, el processament de llenguatge natural, l'extracció i recuperació de la informació o l'aprenentatge automàtic, entre altres, donen suport al *text mining* (mineria de text). De vegades es confon el *text mining* amb la recuperació de la informació (*Information Retrieval* o IR) (Hearst, 1999). Aquesta darrera consisteix en la recuperació automàtica de documents rellevants mitjançant indexacions de textos, classificació, categorització, etc. Normalment s'utilitzen paraules clau per a trobar una pàgina rellevant. En canvi, *text mining* examina un grup de documents i descobreix informació no continguda en cap document individual del grup; en altres paraules, tracta d'obtenir informació sense haver partit de res (Nasukawa i altres, 2001).

Una aplicació molt popular del *text mining* la trobem a Hearst (1999). Swanson intenta extreure informació derivada de reculls de text. Tenint en compte que els experts només poden llegir una petita part del que es publica en el seu camp, en general no s'assabenten dels nous desenvolupaments que se succeeixen en altres camps. Així, Swanson ha demostrat que cadenes d'implicacions causals dins la literatura mèdica poden conduir a hipòtesis per a malalties poc freqüents, algunes de les quals han rebut proves de suport experimental. Investigant les causes de la migranya, l'investigador esmentat va extreure diverses proves a partir de títols d'articles presents en la literatura biomèdica. Algunes d'aquestes claus van ser:

- L'estrès està associat amb la migranya.
- L'estrès pot conduir a la pèrdua de magnesi.
- Els bloquejadors de canals de calci prevenen algunes migranyes.
- El magnesi és un bloquejador natural del canal de calci.
- La depressió cortical disseminada (DCD) té a veure amb algunes migranyes.
- Els nivells alts de magnesi inhibeixen la DCD.
- Els pacients amb migranya tenen una alta agregació plaquetària.
- El magnesi pot suprimir l'agregació plaquetària.

Aquestes claus suggereixen que la deficiència de magnesi podria influir en alguns tipus de migranya, una hipòtesi que no hi havia en la literatura i que Swanson va trobar mitjançant aquells enllaços. D'acord amb Swanson (Swanson i altres, 1994), estudis posteriors han provat experimentalment aquesta hipòtesi obtinguda per *text mining*, amb bons resultats.

5. Conclusions

Títol original: *Data mining: torturando a los datos hasta que confiesen*

La nostra capacitat per a emmagatzemar dades ha crescut en els últims anys a velocitats exponencials. En contrapartida, la nostra capacitat per a processar-les i utilitzar-les no ho ha fet. Per aquest motiu, el *data mining* es presenta com una tecnologia de suport per a explorar, analitzar, comprendre i aplicar el coneixement obtingut manejant grans volums de dades. Descobrir nous camins que ens ajudin en la identificació d'estructures interessants en les dades és una de les tasques fonamentals en el *data mining*.

En l'àmbit comercial, resulta interessant trobar patrons ocults de consum dels clients per a poder explorar nous horitzons. Saber que un vehicle esportiu corre un risc d'accident gairebé igual al d'un vehicle normal quan el seu amo té un segon vehicle a casa ajuda a crear noves estratègies comercials per a aquest grup de clients. Així mateix, prevenir el comportament d'un futur client, basant-se en les dades històriques de clients amb el mateix perfil, ajuda a poder retenir-lo durant el temps màxim possible.

Les eines comercials de *data mining* que hi ha actualment al mercat són variades i excel·lents. N'hi ha que estan orientades a l'estudi del web o a l'anàlisi de documents o de clients de supermercat, mentre que altres són d'ús més general. L'elecció correcta depèn de la necessitat de l'empresa i dels objectius a curt i llarg termini que pretengui assolir. La decisió a l'hora de triar una solució de *data mining* no és una tasca simple. És necessari consultar experts en l'àrea amb vista a seleccionar-ne la més adequada per al problema de l'empresa.

Com s'ha vist al llarg d'aquest article, són moltes les àrees, les tècniques, les estratègies, els tipus de bases de dades i les persones que intervenen en un procés de *data mining*. Els negocis requereixen que les solucions tinguin una integració transparent en un ambient operatiu. Això ens porta a la necessitat d'establir estàndards per a fer un ambient interoperable, eficient i efectiu. Actualment es duen a terme esforços en aquest sentit. A Grossman i altres (2002) s'exposen algunes iniciatives per a aquests estàndards, incloent-hi aspectes en:

- Models: per a representar dades estadístiques i de *data mining*.
- Atributs: per a representar la neteja, la transformació i l'agregació d'atributs utilitzats com a entrada en els models.
- Interfícies i API: per a facilitar la integració amb altres llenguatges o aplicacions de programari i API.
- Configuració: per a representar paràmetres interns requerits per a construir i fer servir els models.
- Processos: per a produir, desplegar i utilitzar models.
- Dades remotes i distribuïdes: per a analitzar i explorar dades remotes i distribuïdes.

En resum, el *data mining* es presenta com una tecnologia emergent, amb diversos avantatges: d'una banda, resulta un bon punt de trobada entre els investigadors i les persones de negocis; d'altra banda, estalvia grans quantitats de diners a una empresa i obre noves oportunitats de negocis. Treballar amb aquesta tecnologia implica, sens dubte, tenir cura d'un gran nombre de detalls, perquè el producte acabat comporta "presa de decisions".

Llista d'URL:

[url1]:http://www.lsi.upc.es/~lcmolina/about.htm

Títol original: *Data mining: torturando a los datos hasta que confiesen*

[url2]:http://www.altavista.com

[url3]:http://www.kdnuggets.com

[url4]:http://www.fcw.com/fcw/articles/2002/0603/news-fbi-06-03-02.asp

[url5]:http://tierra.ucsd.edu/archives/ats-l/2002.06/msg00013.html

[url6]:http://www.fairisaac.com/page.cfm/press_id=325

[url7]:http://www.mining.dk/SPSS/Nyheder/nr7case_bbc.htm

[url8]:http://www.msnbc.com/news/756968.asp

[url9]:http://domino.research.ibm.com/comm/wwwr_thinkresearch.nsf/pages/datamine296.html

Bibliografia:

BRACHMAN, R.J.; KHABAZA, T.; KLOESGEN, W.; PIATETSKY-SHAPIRO, G.; SIMOUDIS, E. (1996). "Mining business databases". *Communications of the ACM*. Vol. 39, pàg. 42-48.

BRODLEY, C.E.; LANE, T.; STOUGH, T.M. (1999). "Knowledge discovery and data mining". *American Scientist*. Vol. 86, pàg. 55-65.

FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (ed.) (1996). *Advances in knowledge and data mining*. Cambridge (Massachusetts): AAAI/MIT Press.

FAYYAD, U.; HAUSSLER, D; STOLORZ, P. (1996). "Mining scientific data". *Communications of the ACM*. Vol. 39, pàg. 51-57.

FELDMAN, R.; DAGAN, I. (1995). "Knowledge discovery intextual databases (KDT)". A: *Proceedings of the 1st international conference on knowledge discovery*. ACM.

GROSSMAN, R. L.; HORNIK, M.F.; MEYER, G. (2002). "Data mining standards initiatives". *Communications of ACM*. Vol. 45 (8), pàg. 59-61.

HEARST, M. (1999). "Untangling text data mining". A: *Proceedings of 37 th annual meeting of the association for computational linguistics*. Universitat de Maryland.

KOSALA, R.; BLOCKEEL, B. (2000). "Web mining research: a survey". *SIGKDD Explorations: Newsletter of the special interest group on knowledge discovery and data mining*. ACM Press. Vol. 2 (1).

MENA, J. (1999). *Data mining your website*. Digital Press.

MOLINA, L.C. (1998). *Data mining no processo d'extração de conhecimento de bases de daus*. Tesi de màster. Sâo Carlos (Brasil): Institut de Ciências Matemáticas i Computação. Universitat de São Paulo.

MOLINA, L.C.; RIBEIRO, S. (2001). "Descubrimiento conocimiento para el mejoramiento bovino usando técnicas de Data Mining". A: *Actes del IV Congrés Català d'Intel·ligència Artificial*. Barcelona, pàg. 123-130.

NASUKAWA, T.; NAGANO, T. (2001). "Text analysis and knowledge mining system". *IBM Systems Journal, knowledge management*. Vol. 40 (4).

RODAS, J. (2001). "Un ejercicio de análisis utilizando Rough Sets en un dominio de educación superior mediante el proceso KDD". Document intern. Barcelona: Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya.

SWANSON, D.R.; SMALHAISER, N.R. (1994). "Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease". *Neuroscience research communications*. Vol. 15, pàg. 1-9.

WAY, J.I.; SMITH, E.A. (1991). "The evolution of synthetic aperture radar systems and their progression to the EOS SAR". *IEEE Transactions on geoscience and remalnom sensing*. Vol. 29 (6), pàg. 962-985.

ZAIANE, O.R.; HAN, J.; LI, Z-N.; CHEE, S. H.; CHIANG, J.Y. (1998). "MultiMedia-Miner: a system prototype for multimedia data mining". A: *Proceedings of the international conference on management of data*. ACM SIGMOD. Vol. 27 (2), pàg. 581-583.

Títol original: *Data mining: torturando a los datos hasta que confiesen*

Enllaços relacionats:

- Formació a la UOC
  http://www.uoc.edu/masters/cat/cursos/especialitzacio/208_ct.html
- KDnuggets
  http://www.kdnuggets.com/
- KDcentral
  http://www.kdcentral.com/
- *Data Mining and Knowledge Discovery. An International Journal*
  http://www.digimine.com/usama/datamine/
- Departament de Llenguatges i Sistemes Informàtics

Grup de Soft Computing
  http://www.lsi.upc.es/~webia/soft-comp.html

- Pàgina de Luis Carlos Molina Félix
  http://www.lsi.upc.es/~lcmolina/

Títol original: *Data mining: torturando a los datos hasta que confiesen*