1 Índex

2 Introducció

2.1 Motivació i objectius

L'objectiu d'aquest projecte és l'estudi de la integració de mètodes de mineria de dades en sistemes de gestió de bases de dades, concretament en els de lliure distribució. Es pretén aprendre tant conceptes generals de la mineria de dades com propis d'algun algortime concret. S'aprofunditzarà en les possibilitats d'un sistema de gestió de bases de dades de lliure distribució, i s'estudiaran vàries de les facetes vinculades tant a la instal.lació, com per exemple els sistemes operatius de lliure distribució, com al desenvolupament, amb entorns de diferent tipologia.

Personal i professionalment, el tracatament de dades, informació i coneixement és un camp que desperta molta curiositat. Optimització i eficiència són conceptes que sempre s'han d'utilitzar a l'hora de dissenyar aplicacions d'enginyeria, així que s'en farà un especial èmfasi al llarg del projecte.

2.2 Organització de la memòria

En una primera part es farà una introducció general a la mineria de dades, contextualitzant-la a la informàtica actual, i seguidament es donarà pas a una descripció tant del procés de coneixement com del pas que ens interessa, la mineria de dades. Un dels punts on es profunditzarà sobre la mineria de dades és la seva relació amb els sistemes gestors de bases de dades, i s'avaluaran els pros i els contres de la integració de les funcionalitats.

Posteriorment s'entrarà en l'estudi d'un del mètodes de mineria de dades més populars, les regles associatives, concretant tres possibles implementacions integrades de l'algoritme a la base de dades de Postgres.

Les decisions preses en temps de disseny i programació seran analitzades empíricament, i amb els resultats d'aquestes obtindrem una sèrie de reflexions interessants per a l'orientació de futurs desenvolupaments.

3 Mineria de dades

3.1 Contextualització històrica

La informàtica ha tingut sempre com a un dels seus principals objectius el tractament de la informació, i a mida que la tecnologia ha anat evolucionant, la informació o el coneixement que demanem a la informàtica també ha evolucionat.

Així doncs, en el món dels negocis als anys 60, quan la tecnologia posava l'èmfasi en l'emmagatzematge de les dades, el que es podia obtenir de la informació era, per exemple, la resposta a "Quin és el total de vendes de productes en els últims 2 anys?", ja que els computadors de l'època i la tecnologia de cintes només permetien l'anàlisi retrospectiu de dades estàtiques.

A partir dels anys 80, on preocupava l'accés a les dades, van aparèixer les bases de dades relacionals i un llenguatge estàndar per a fer consultes (SQL), la informació demandada era molt més precisa, a nivell de registre, i aleshores la pregunta a resoldre es podia transformar en "Quantes unitats he venut de CocaCola a l'abril?".

Als anys 90, amb les bases de dades orientades a l'anàlisi, i a mida que la tecnologia avançava amb les bases de dades multidimiensionals i el procés analític, la pregunta va evolucionar a "Quines unitats de CocaCola he venut a Catalunya l'últim abril, i quines he venut a nivell de Barcelona?". Així doncs, ens trobàvem ja amb informació multinivell o multidimensional.

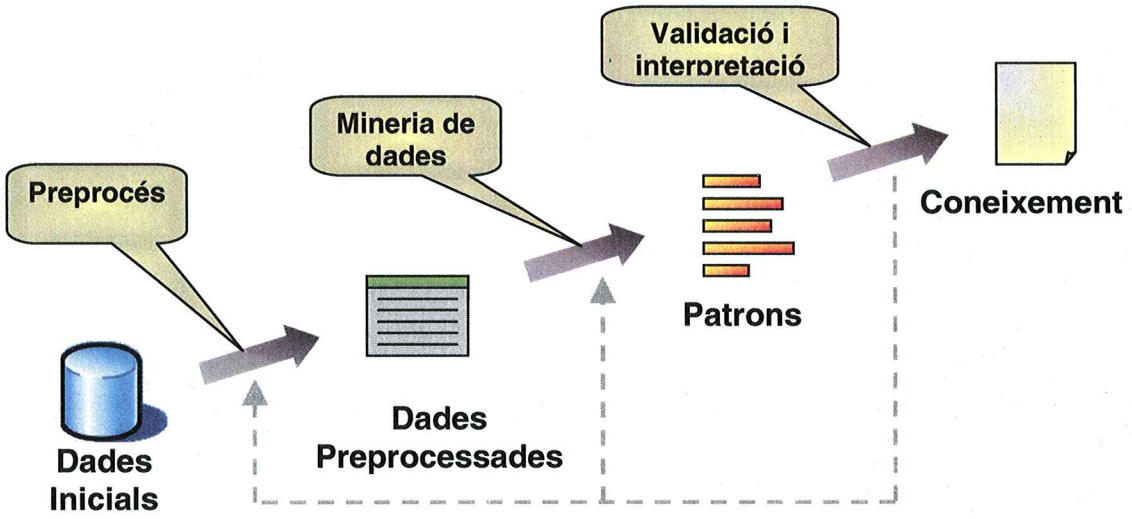
Actualment, amb les tècniques de cerca i els models predictius utitlitzats en busca de coneixement amb la mineria de dades, i amb la utilització d'algoritmes avançats i servidors multiprocessador, la pregunta a resoldre es tranforma en "Què em compraran aquests clients el pròxim mes?"

Aquesta evolució de la demanda ha estat acompanyada, a més, per un creixement de la informació emmagatzemada a analitzar, tant en el nombre de files com en el nombre de columnes, de manera que el tractament d'aquest gran volum d'informació ha quedat fora de l'abast de la capacitat humana per a poder-la processar i analitzar.

L'objectiu és utilitzar la tecnologia per aconseguir extreure, de grans volums de dades, informació sintetitzada i més fàcil d'assimilar. Aquesta informació que podrem extreure serà de tipus quantitatiu, per exemple un informe de resultats numèrics, o descriptiu, com la d'un model predictiu.
Cal destacar també que aquesta necessitat de la informació neix del món "real", i no de l'acadèmic, de manera que seran moltes les aplicacions i diversos els camps on podrem utilitzar la tècnica de minera de dades.

3.2 Procés d'obtenció del coneixement

La mineria de dades és un dels passos dins del procés d'obtenció de coneixement. Aquest conjunt de passos es poden agrupar dins de tres grups principals: el preprocés de les dades, la pròpia mineria de dades i la validació i interpretació dels patrons.



En primer lloc, ens cal analitzar quin és el domini aplicatiu del coneixement que tenim com a objectiu, per tal de poder crear un conjunt de dades adequat a aquest objectiu, i així reduir les variables a tenir en compte .

En un del exemples de mineria de dades anomenat "la cistella de la compra" on es busquen les relacions entre articles d'una compra, podem decidir prescindir d'informacions com són l'hora de la compra o la persona que ha atès el client, ja que considerem que no són variables rellevants.

També pot ser necessari acotar o reduir el conjunt de dades a analitzar en funció de l'objectiu final de la tasca. És a dir, si el que ens interessa saber és com ordenar els productes en un supermercat durant els pròxims 2 mesos, caldrà analitzar les dades de les compres realitzades en anys anteriors durant només aquests 2 mesos, ja que en les vendes de productes és rellevant l'època de l'any.

Hem de tenir en compte, també, que abans de la utilització de les tècniques de mineria de dades, acostuma a ser necessari seguir un conjunt de passos per a poder depurar i transformar les dades.

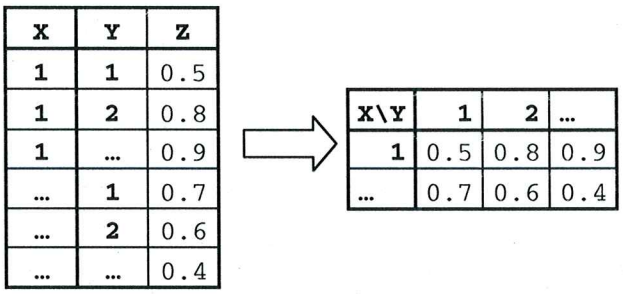
La depuració és necessària ja que sovint, en les dades a tractar, hi ha valors fora dels paràmetres que volem estudiar, de manera que ens caldrà netejar les dades per tal d'eliminar al màxim el soroll i els possibles errors, i així els resultats del nostre procés seran més ajustats a la realitat.

Per exemple, en l'aplicació de "la cistella de la compra" podríem eliminar tota la informació relativa a les compres d'un sol article, ja que encara que no aportés cap tipus de regla, ens influenciaria en els càlculs probabilístics de les altres regles.

D'altra banda, els sistemes d'informació no sempre tenen les dades en el format que necessiten d'entrada les tècniques de mineria de dades, de manera que caldrà processar les dades per deixar-les en un model que ens serveixi després per a poder aplicar la mineria de dades.

Sovint, quan es plantegen els models de dades de les aplicacions, ja es té en compte si s'en farà aquest ús. Si és així, ja es tindrà la informació en el model adequat.
Les tècniques de procés analític online ja ens permeten fer aproximacions al model desitjat gràcies al tractament multidimensional de les dades.
Un exemple de tractament de la informació seria quan tenim un algoritme que requereix que les dades estiguin en format de matriu, però en el model relacional estan emmagatzemades com a files de poques columnes.

Així doncs, caldria fer la següent transformació:



X	Y	Z
1	1	0.5
1	2	0.8
1	...	0.9
...	1	0.7
...	2	0.6
...	...	0.4

X\Y	1	2	...
1	0.5	0.8	0.9
...	0.7	0.6	0.4

Cal també escollir correctament els objectius del procés de coneixement, ja que amb aquests objectius i la selecció d'hipòtesis triarem el mètode de mineria de dades apropiat per al procés.

Un cop es té la informació a processar i en el format que convé, el següent pas és el de la minera de dades, que mitjançant mètodes algorítmics processarà aquest gran volum de dades per reduïr-les a un model més manejable. Aquest pas és el nostre objecte d'estudi, pel que més endavant entrarem en més detall.

Després de la mineria de dades només ens quedarà donar un últim pas, la interpretació dels resultats, ja sigui extraient les conclusions des d'un model descriptiu, com extraient-les des d'un model predictiu.

Un exemple de representació d'un model descriptiu podria ser un informe. Per un model predictiu, en canvi, si bé la representació del model és important per a la interpretació, és més important l'ús que s'en faci, ja que típicament el model s'acabarà implementant en alguna aplicació informàtica que permeti l'explotació del model predictiu.

Amb l'exemple de "la cistella de la compra", les regles obtingudes podrien ser les següents:

Bolquers → Cervesa
Pa ∧ Tomàquet → Oli
...

Aquestes regles s'utilitzarien en una aplicació informàtica que generés distribucions dels productes dins d'un supermercat, de manera que es maximitzessin les compres.

Aquest procés de l'obtenció del coneixement es pot repetir de manera iterativa, ja que amb les conclusions obtingudes es poden replantejar algunes de les decisions en quant a tècniques, variables o objectius que s'han pres durant el procés.

Les aplicacions d'aquests processos són molt diverses. En Màrketing, per exemple, una de les aplicacions més habituals de la mineria de dades és utilitzar-la per a obtindre classificacions de grups de consumidors, i així poder millorar els objectius de publicitat, o la coneguda aplicació de "la cistella de la compra", la qual intenta deduir patrons de compra, com per exemple "Si algú compra X i Y, probablement comprarà Z".

Un altre exemple d'aplicació pràctica de la minera de dades el podem trobar en la detecció del frau, on s'apliquen sistemes per monotoritzar les milions d'operacions fetes amb targetes de crèdit, i detectar així conductes típicament delictives. Com també s'han utilitzat tècniques de mineria de dades per identificar transaccions financeres que indiquin el blanqueig de diners.

En el món de la ciència també s'utilitzen sovint algoritmes de mineria de dades per a la detecció de patrons en laboratoris d'investigació, com també s'utilitzen per a diagnosi de malalties en medicina.

3.3 Fonaments de la Mineria de dades

Amb el creixement actual de la informació, tant a nivell d'unitats d'informació (registres) com a nivell de variables (columnes), s'han elaborat mètodes digitals de tractament d'aquesta informació per aconseguir-ne el seu coneixement, ja que queda fora de l'abast de la capacitat humana d'aconseguir-ho de forma manual.

Dins del procés per descobrir el coneixement, la mineria de dades és el pas consistent en l'aplicació d'anàlisis de les dades i algoritmes de descobriment, que sota unes acceptables limitacions d'eficiència computacional, produeixin una enumeració particular de patrons o models sobre les dades.

D'altra banda, el fet de que no sigui una disciplina rígida i de que hi hagi un nombre considerable de problemes diferents i de tècniques per resoldre'ls, ressalta la necessitat de fer un esforç incial per concretar l'objectiu de la cerca del coneixement, ja que l'objectiu ens marcarà el mètode a emprar.

Així doncs, podem dividir els objectius en predictius o en descriptius. En els primers s'utilitzen variables i camps de la base de dades per crear un model del que es puguin deduir futurs valors d'interès. En els descriptius la cerca es centra en trobar patrons interpretables per a l'home que descriguin les dades. Sovint, els límits entre uns i altres no estan perfectament definits, de manera que la relativa importància de la predicció o la descripció per a una particular aplicació de mineria de dades pot variar considerablement.

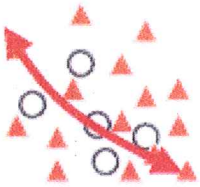
Els objectius de la predicció i la descripció es poden abordar utilitzant una varietat particular de mètodes de mineria de dades.

3.4 Tècniques

Regressió

L'objectiu del mètode regressiu és buscar una funció de mapeig d'una dada real que ens permeti predir una variable.

Per exemple, preveure la probabilitat de que un pacient pugui sobreviure donats els resultats d'un conjunt de tests, o la demanda que tindrà un consumidor davant un nou producte al veure un anunci, tenint en compte el comportament dels consumidors en experiències anteriors.

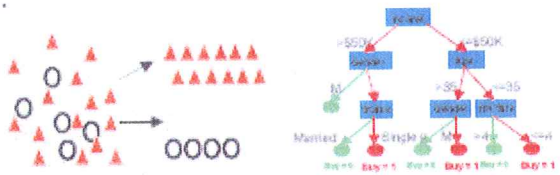


Classificació

Es busca una funció que permeti catalogar un element de dades en una de les classes predefinides.

Un dels exemples d'aquest mètode és la classificació de textos en idiomes de forma automatitzada, o el descobriment de tendències a l'alça o a la baixa en inversions financeres.

Cal destacar que la funció que defineix la divisió que es fa dels elements pot ser imperfecte, de manera que hi hagi elements que no segueixin aquest patró.

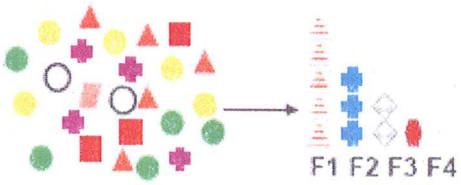


Sumarització

Fa referència a mètodes per a la busca d'una descripció compacta d'un subconjunt de dades.

Un exemple seria la tabulació de les desviacions estàndar de tots els camps d'una relació d'elements.

Aquestes tècniques sovint són aplicades a l'exploració interactiva de dades o a la generació d'informes automatitzats.

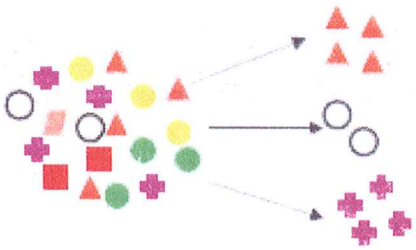


Clustering (Agrupació)

És una tasca comunament descriptiva amb la qual s'identifica un conjunt finit de categories o grups per descriure les dades. Les categories poden ser exclusives i exhaustives, o consistir en una representació d'herència o continència.

Una de les aplicacions més habituals d'aquest mètode és l'agrupació dels consumidors en categories homogènies, amb la possiblitat de la creació de subcategories.

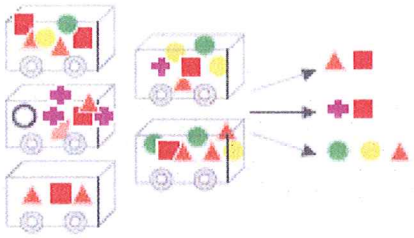
Es basa amb l'estimació de probabilitats de densitat, és a dir, en agrupar els elements que estan més a prop tenint en compte les variables analitzades.



Models de dependència

Consisteix en la cerca d'un model que descrigui les dependències significatives entre variables, ja sigui a nivell estructural, on les variables depenen unes de les altres, o a nivel quantitatiu, que defineix la força de les dependències sobre una escala numèrica.

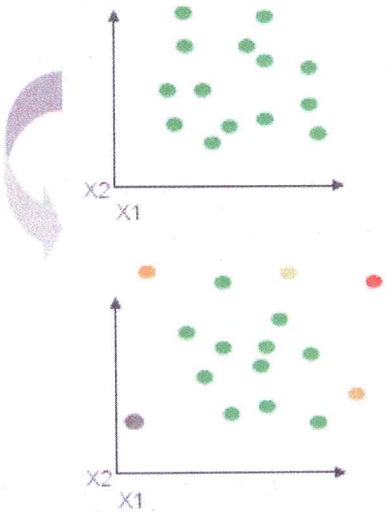
Algunes de les aplicacions d'aquest mètode són el desenvolupament de sistemes mèdics experts, recuperació d'informació, i modelització del genoma humà.



Detecció de canvi i desviació

Es focalitza en el descobriment dels canvis més significatius en dades prèviament mesurades i en normatives.

S'han utilitzat tècniques com aquesta per a la detecció dels fenomens relacionats amb el canvi climàtic.



3.5 Components dels algoritmes de mineria de dades

S'identifiquen 3 components en qualsevol implementació d'un algoritme de mineria de dades:

La representació del model, que és el llenguatge utilitzat per descriure els patrons descoberts.
És important que sigui suficientment extens per representar les suposicions i decisions que s'han pres, així com que sigui comprensible i viable per al testeig.

Els criteris d'avaluació del model, que són sentències quantitatives o funcions de fita per als patrons particulars obtinguts en el procés d'obtenció de coneixement.
Aquests criteris ens marcaran la precisió dels patrons obtinguts, com també han de ser introduits per motius d'eficiència computacional.

El mètode de cerca, format pels paràmetres de la busca, els quals fixen els criteris d'avaluació, i pel model de la busca, sovint representat com un bucle on van variant els paràmetres de cerca del mètode. Aquest algoritme retornarà un model predictiu, generalment en forma de conjunt de patrons o regles.

De cada model de cerca s'en poden obtindre una varietat d'algoritmes i tècniques, que cada un dins el seu context, tenen la seva representació del model i els seus criteris d'avaluació.

Per popularitat destaquen els següents:

Arbres de Decisió i Regles

Són una forma de representació de divisions amb un model relativament fàcil de comprendre. Les divisions representen restriccions significatives d'un model.

Regressions no lineals i mètodes de classificació

Són una família de tècniques de predicció que plantegen convinacions de funcions lineals i no lineals amb les variables d'entrada. Com més complicada és la funció, més precisa és, però també més difícilment interpretable.

Basats en exemples

La representació és molt simple, utilitza exemples de la base de coneixement per aproximar el model, així les prediccions es basen en les similituds de propietats d'altres exemples.
Aquestes tècniques inclouen mètodes com el nearest-neighbor (veïns més propers).
L'inconvenient del mètode és que cal una molt bona definició de la distància en els criteris d'avaluació.

Models gràfics de dependències probabilístiques

Aquestes tècniques utilitzen l'estructura de graf, que és una estructura simple on el model especifica quines variables són directament dependents d'una altra variable.
Típicament aquests models s'utilitzen en sistemes exeperts en intel.ligència artificial.

Models relacionals d'aprenentatge

Són una representació restringida de lògica proposicional, és a dir, lògica inductiva de programació, que utilitza el patró flexible de lògica de primer ordre.
Es basa en la facilitat d'interpretació i la simplicitat del llenguatge a l'hora de satisfer les demandes computacionals dels sistemes de cerca.

3.6 Consideracions

Les bases de dades que són objecte de la mineria de dades sovint són de grans dimensions, no tan sols amb centenars de milions de registres, sinó també amb una gran quantitat d'atributs i variables. Aquest fet comporta la inevitable necessitat d'optimitzar tant l'accés a les dades com el procés que es faci amb elles.

D'altra banda, el fet de que hi hagi moltes variables també farà que l'espai de busca creixi, ja que en la busca de les relacions entre variables es crearan noves dimensions d'anàlisi.
No tan sols es genera una gran quantitat d'informació, sinó que es genera de manera molt ràpida, i el procés d'obtenir coneixement d'aquesta informació ha de ser prou eficient com perquè no es generi un model obsolet.

Quan amb un algoritme es busca en un subconjunt de dades, utilitzant el que creiem els millors paràmetres i criteris de poda, i amb algun model particular, se n'obtenen uns patrons. És possible que el soroll del subconjunt de dades faci que el resultat no sigui generalista, així doncs caldrà fer validacions creuades, regularitzacions i altres estratègies estadístiques per solucionar el problema de la fiabilitat.

Un problema relacionat amb l'anterior és l'elecció de variables estadístiques significatives, caldrà doncs tenir-ho en comtpe durant el procés de coneixement, ja que possiblement s'hauran de reconsiderar valors i repetir passos del procés.

Si bé és important l'eliminació del soroll en un conjunt de dades a analitzar, a la pràctica no és un pas senzill, ja que requereix un coneixement del model resultant abans d'executar el procés. Per aquest motiu el procés d'obtenció del coneixement ha de ser un procés interactiu, en el qual qui dirigeix el procés ajusti les decisions, i també iteratiu, per replantejar-se algun dels passos del procés i repetir-lo amb uns nous criteris d'avaluació del model.

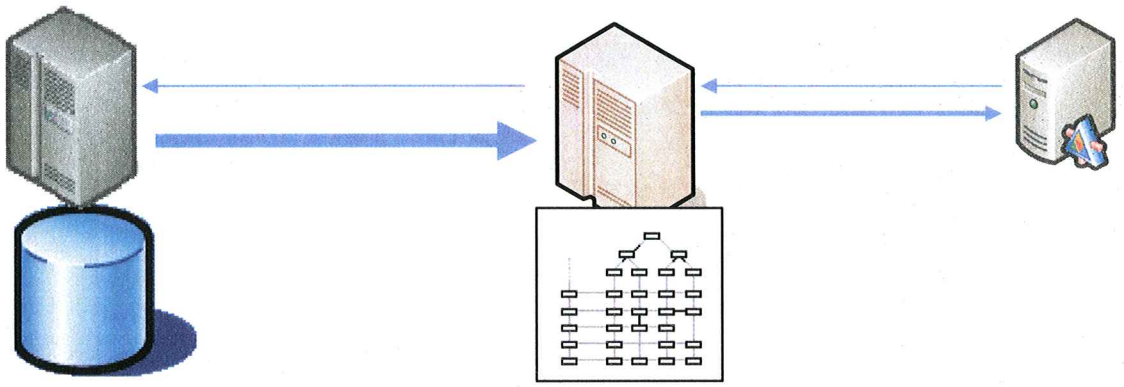
En un primer moment, les tècniques de mineria de dades s'implementaven en aplicacions independents de l'accés a les dades, així doncs el model que s'utilitzava típicament era el d'una aplicació client que recuperava les dades d'un servidor de base de dades.

Degut al gran volum de dades que es mouen a cada iteració el model no era prou eficient, d'aquesta manera neix la necessitat d'integrar l'accés a les dades i el tractament d'aquestes el més possible per tal d'optimitzar el procés.

4 Bases de dades i mineria de dades

4.1 Problemàtica de tenir la mineria de dades fora de la base de dades

En un model d'aplicació de mineria de dades on el procés de les dades no està integrat amb l'accés a aquestes, ens trobem amb un conjunt de dificultats computacionals afegides al ja complex procés d'obtenció del coneixement.



En aquesta arquitectura el servidor de càlcul i anàlisi, un cop rep la instrucció de començar el procés, solicita al gestor de base de dades totes les dades que hauran de ser recuperades del disc pel gestor i transmeses al servidor de càlcul.

Evidentment, aquesta recuperació de dades no és necessari que es dugui a terme d'una sola vegada, ja que pot executar-se per parts.
Com que la lògica funcional de l'aplicació està al servidor que executa l'algoritme, el sistema gestor de base de dades no ho tindrà fàcil per a utilitzar la seva potència d'optimització.
Així, el procés es penalitza l'eficiència amb la recuperació i transmisió de les dades.

A mida que les dades arriben al servidor de càlcul, l'algoritme va creant una complexa estructura de dades a memòria.

Cal destacar que el fet de que es crei una estructura a mida per l'algoritme fa créixer l'eficiència del procés. Però tenint en compte la naturalesa de la mineria de dades, on un dels trets més caraterístics és el gran volum de dades, el fet d'intentar emmagatzemar un volum de dades memòria més gran que la capacitat del servidor farà que part d'aquesta informació s'hagi d'emmagatzemar a disc i utilitzar tècniques d'intercanvi de memòria

(swapping) que reduiran l'eficiència d'aquesta estructura a memòria, i per tant, del procés.

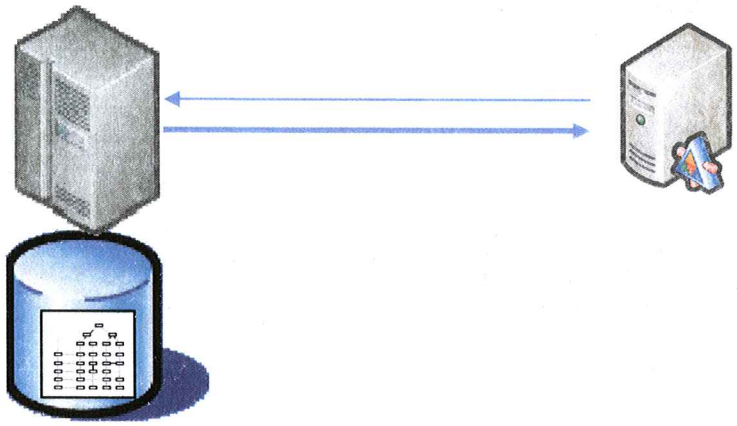
La busca d'aquesta estructura òptima de memòria, de l'algoritme de recorregut de l'estructura i del procés de les dades pot arribar a ser molt costosa, i amb el coneixement obtingut a posteriori del procés es podria millorar a cada iteració, amb l'alt cost d'inversió que suposaria.

D'altra banda, els sistemes gestor de bases de dades ja incorporen avançades tècniques de creuament de dades, que amb comandes d'alt nivell són molt més fàcils d'administrar.

Finalment, el conjunt de resultats serà retornat a l'usuari, ara bé, a menys que sigui d'utilització puntual, hauran de ser emmagatzemats en una base de dades per a la seva posterior utilització, així que un nou volum de dades viatjarà del servidor de càlcul a un sistema d'emmagatzematge.

També cal destacar com a aspecte positiu d'un model distribuit, que amb l'especialització dels servidors es poden obtindre millores considerables en l'eficiència del procés.

En un model integrat, on el sistema gestor de base de dates té implementats algoritmes de miniera de dades, l'usuari enviarà comandes d'alt nivell al servidor per arrancar el procés.



El sistema gestor de base de dades rep la comanda d'executar l'algoritme de mineria de dades. Aquest algoritme ha d'estar dissenyat tenint en compte les possibilitats d'optimització en el tractament de dades que tenen els gestors.

Els actuals gestors de bases de dades ens informen de quin és el pla d'execució dels tractaments de dades, per exemple creuaments, i els més comercialitzats fins i tot poden sugerir la creació dels índexs que optimitzin aquests procesos.

Així doncs, l'existència d'aquests poderosos sistemes d'optimització és una gran avantatge que necessàriament s'ha d'explotar.

Per una altra banda, com a sistemes gestors de dades, tenen optimitzacions a l'hora d'accedir a les dades de disc, i permeten configuracions en aquest sentit que poden ajudar a l'eficiència del procés.

Un altre avantatge respecte a un sistema no integrat és que les dades obtingudes de disc no han de ser transferides a un altre sistema, ja que és el propi servidor el que tractarà les dades. El sistema gestor de base de dades podrà llegir-les progressivament de disc a mida que les necessiti, paral.lelitzant així el procés de tractament i de lectura, i estalviant, fins i tot, algún accés a disc.

Finalment, cal destacar que els resultats quedarien emmagatzemats en el propi servidor, pel que novament s'optimitza en el fet de no transportar gran quantitat d'informació.

4.2 Integració de la mineria de dades a les bases de dades

A nivell comercial hi ha dos grans sistemes de gestió de base de dades líders, Oracle i SQLServer, a part del DB2.

Aquests dos sistemes ja incorporen nombroses tècniques de mineria de dades, i en ambdós casos la integració és similar.

Es tracta d'enriquir el llenguatge SQL amb un conjunt de comandes per a processar les dades, però sense seguir un estàndar, ja que cada una de les aplicacions anomena i parametritza les funcions de mineria de dades de forma diferent.

Profunditzant més en l'eina, Oracle ofereix varies de les tècniques de les que hem parlat anteriorment, com seria la detecció d'anomalies, importància d'atributs, regles associatives, agrupament (clustering), classificació i regressió, aplicant diversos algoritmes per a resoldre poblemes de diferent naturalesa.

Problema	Algoritme	Aplicabilitat
Classificació	**Arbres de decisió** **Naïve Bayes** **Support Vector Machine** **Adaptive Bayes Network**	Popular/Rules/Transparency Embedded app Wide/ Narrow data Regles / Transparency
Regressió	**Support Vector Machine**	Wide / Narrow data
Importància d'atributs	**Descripció Minimum**	Reducció d'atributs Identify useful data Reducció soroll de les dades
Regles d'associació	**A priori**	Anàlisi cistella de la compra Anàlisi de links
Agrupament	**Hierarchical k-means** **Hierarchical o-cluster**	Agrupació de producte Anàlisi de gens i proteïnes
Extracció d'atributs	**NMF**	Anàlisi de dades Reducció d'atributs

Un exemple de la sintaxi d'Oracle convinant SQL estàndar amb les seves comandes integrades de mineria per obtenir dades del llançament d'una campanya seria:

1.Donada una construcció prèvia al model de resposta,...
predir qui respondrà a la campanya,
... i per què
2....investigar
quant gasta cada client tres mesos abans i després de la campanya
3....quant gasta
en DVD's?
4. És l'èxit estadísticament significatiu?

```
select responder, cust_region, count(*) as cnt,
       sum(post_purch – pre_purch) as tot_increase,
       avg(post_purch – pre_purch) as avg_increase,
       stats_t_test_paired(pre_purch, post_purch) as
 significance
from (
 select cust_name,
       prediction(campaign_model using *) as responder,
       sum(case when purchase_date < 15-Apr-2005 then
            purchase_amt else 0 end) as pre_purch,
       sum(case when purchase_date >= 15-Apr-2005 then
            purchase_amt else 0 end) as post_purch
from customers, sales, products&PRODDB
where sales.cust_id = customers.cust_id
  and purchase_date between 15-Jan-2005 and 14-Jul-2005
   and sales.prod_id = products.prod_id
   and contains(prod_description, 'DVD') > 0
 group by cust_id, prediction(campaign_model using *) )
group by rollup responder, cust_region order by 4 desc;
```

En aquest exemple podem veure a la banda esquerra la problemàtica de negoci, i a la banda dreta, en el mateix color que la pregunta corresponent, com es traduiria a llenguatge SQL enriquit amb funcions de miniera de dades.

El sistema gestor de base de dades de Microsoft és el SQLServer, que en la seva versió 2005 ja incorpora forces comandes de mineria de dades, també integrades dins del seu llenguatge processual SQL enriquit.

Alguns dels algoritmes que implementa són els arbres de decisió, regles associatives, Native Bayes, Agupació, Sèries temporals, Xarxes neuronals o busca en textos desestructurats.

La utilització d'aquestes tècniques a les consultes SQL estàndar és força amigable. Per exemple:

```
SELECT TOP 25 t.CustomerID FROM CustomerChurnModel NATURAL
PREDICTION JOIN OPENQUERY('CustomerDataSource', 'SELECT * FROM
Customers') ORDER BY PredictProbability([Churned],True) DESC
```

Analitzant ara dos dels gestors més populars de lliure distribució, el MySQL i el PostgreSQL, trobem que tot i tenir l'avantage de que són gratuïts, també responen amb un bon rendiment respecte l'accés a les dades. Ara bé, com a contrapartida, la inversió en la millora i escalabilitat del producte no està controlada i orientada com en els productes comercials. Així doncs, la seva extensió depèn de les aportacions fetes pels seus usuaris desenvolupadors, d'aquesta manera l'aplicació va creixent i evolucionant.

Un dels objectius d'aquest projecte és analitzar les possiblitats que ens ofereix Postgres en quan a creixement en el món de la mineria de dades. Així doncs, caldrà profunditzar en l'optimització dels procesos de desenvolupament i explotació.