

Introduction à l'extraction de connaissances et à la fouille de données

HLIN208

Pascal Poncelet
LIRMM
Pascal.Poncelet@lirmm.fr
<http://www.lirmm.fr/~poncelet>



Extraction de connaissances

- (une petite parenthèse avant de commencer)



2

Quelques chiffres

- 2,3 milliards dans le monde
 - 41,2 millions en France
 - 32 millions sont inscrits sur au moins un réseau social
 - 52 % ont entre 25 et 45 ans
 - Facebook : 26 000 000 utilisateurs Français sur 1,6 milliards de membres !!
 - En moyenne les utilisateurs ont 177 amis :)

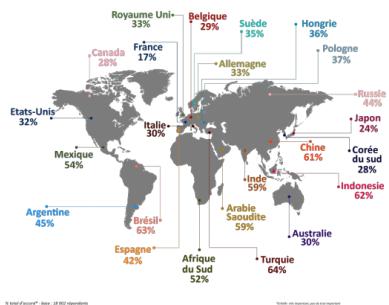
(Sources Factory.com (2013))



iii

De l'utilisation des sites de réseaux sociaux

Ipsos OTX
Open Thinking Exchange
« Les réseaux sociaux sont très/assez importants pour moi ? »



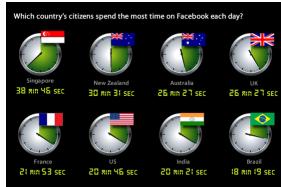
% total d'accès** - Date : 18.02.2012 répondants

**Méthode : 1000 répondants par pays et par sexe

4

Beaucoup de temps

- 66 % des utilisateurs Français se connectent une fois par jour
- 22,5 % du temps de connexion est associé aux réseaux sociaux



ATTENTION :

« Un homme de 42 ans employé dans une entreprise du Maine-et-Loire vient d'être congédié pour un usage abusif de Facebook sur son lieu de travail.

La même sanction avait été retenue début septembre contre une salariée des Pyrénées-Atlantiques. »

5

De l'utilisation des sites de réseaux sociaux

- Sondage mené par Harris Interactive,
 - 45% des recruteurs Américains déclarent utiliser les sites de réseaux sociaux (Facebook, MySpace, LinkedIn, Twitter, etc.) pour trouver des informations sur des candidats qui postulent à leurs offres d'emploi
 - 35% ont écarté des candidats en raisons ce que qu'ils ont trouvé :
 - 53 % publication du candidat de photos ou d'informations provocantes ou déplacées
 - 44 % parce que l'on voit les candidats buvant ou se droguant
 - 35 % parce qu'ils crachaient sur leurs anciens employeurs, leurs collègues ou leurs clients
 - 29 % parce qu'ils montraient un déficit de communication
 - 26 % parce qu'ils publiaient des propos discriminatoires
 - 24 % parce qu'ils mentaient sur leurs diplômes et
 - 20 % parce qu'ils ont publié des informations confidentielles sur leurs anciens employeurs
 - Allemagne : 28% des employeurs (500 entreprises) utilisent Internet pour recueillir des informations dès le début du recrutement

6

Les amis de mes amis

- Entretien avec Alex Türk, président de la Cnil (Commission nationale de l'informatique et des libertés).
 - « Un de ses copains a pris la photo et l'a balancé sur le réseau social. C'est amusant. Quelques mois plus tard, il était candidat sur un poste et le recruteur lui a glissé sous les yeux la photo de ses fesses en lui demandant s'il était coutumier de ces pratiques ». Source ([site Internet du quotidien La Provence](#))

□ « Oh mon dieu ! Je hais mon boulot » ajoutant que son responsable était « pervers » et qu'il ne lui donnait que « du travail de m... »

□ ...4 heures plus tard...

□ « Tout d'abord arr



Notre responsabilité

- Expérience de l'éditeur britannique Sophos (2007)
 - Création d'un compte Freddy Staur
 - Envoi de Friends à un échantillon de 200 personnes sur FaceBook
 - **87** personnes ont répondu en donnant accès à des photos de familles, des informations sur leur goûts, le nom de leur compagnon, compagne, (le nom de jeune fille de leur mère) leur CV



8

Une expérience

- Take this lollipop:
 - <http://www.youtube.com/watch?v=SnAxsXOcrkw>
 - Vous pouvez essayer :
 - <http://www.takethislollipop.com/>
 - Attention : vous donnez votre adresse facebook ☺
êtes vous sur ?



10

Les moteurs de recherche

- Les photos de Laure Manaudou - décembre 2007

Google Trends search history for 'laure manaudou carla bruni' from Nov 29 2007 to Dec 16 2007. The graph shows a sharp increase in interest starting around December 10th. Regions showing high interest include France, Toulouse, France, Bordeaux, France, Rennes, France, and Nantes, France.

10

Difficile ?

What does G2P do?
G2P (Google to Peer) uses some crafty Google searches to help locate open directories or otherwise shared files. These searches are nothing weird in fact, take a look at the results of a search for 'dial 999' and you will see that it's much easier to remember .G2P.org than these complex searches. Really? I just think we oughter to make it easier on me, and then share it with you!

Why use G2P instead of P2P or BT?
P2P-BT is being monitored - Using Google we can download it more safely. We are simply just following a link - curious how it looks directly to the file we are looking for?

11

Non - google requêtes complexes

La requête google :

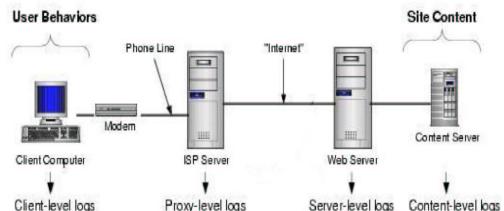
```
intitle:index.of +"Last modified "
+"Parent directory " +(XXXXXXXXX)
+(jpeg) +"" -htm -html -php -asp
```

[XXXXMyBestFriendsXXXXXX.jpg](#)

12

Log ou Logs ?

Information sur les chemins de navigation dans les fichiers logs



13

Web logs + ?

IP or domain name	User Id	Date and Time	Request
123.456.78.9	--	[24/Oct/1999:19:13:44 -0400]	"GET /Images/tagline.gif HTTP/1.0"
200 1449	http://www.teced.com	"Mozilla/4.51 [en] (Win98;!)"	
File Size		Browser	Cookies

[Référer URL](#)

Bases de données des achats
Bases de données des partenaires
Géolocalisation
Cookies



14

Les exemples de clauses

2. COLLECTE ET UTILISATION DES INFORMATIONS

a. *Informations collectées ou reçues de votre part (ou de votre enfant autorisé).*

Nos principaux objectifs dans la collecte d'informations consistent à fournir et à améliorer nos Services, afin d'administrer votre utilisation (ou celle de votre enfant autorisé) et vous permettre (ou à votre enfant autorisé) d'en profiter et d'y naviguer.

i. Informations relatives aux comptes

Informations relatives aux comptes.

endant le jeu et lorsque vous (ou votre enfant autorisé) vous inscrivez pour créer un compte sur nos Services (« Compte »), nous recueillons certaines informations qui peuvent être ensuite utilisées pour vous identifier ou vous reconnaître (ou votre enfant autorisé) (« Données à caractère personnel »). Plus précisément, du fait que vous devez posséder un compte avec Google, « PTC », ou Facebook au moment de la création de votre Compte, nous recevons des informations relatives à ces comptes.

recueillerons les données à caractère personnel (telles que votre adresse e-mail Google, votre adresse e-mail enregistrée sur PTC, et / ou votre adresse e-mail enregistrée sur Facebook) que vous paramétrez de confidentialité sélectionnées sur Google, PTC ou Facebook nous autorisent à accéder.

Lors de l'enregistrement d'un compte PTC, la date de naissance de l'utilisateur et/ou nom d'utilisateur PTC seront demandés (vous concernant ou votre enfant autorisé). Ces informations nous seront partagées (voir le paragraphe « Comptes pour les enfants » dans ce qui suit à la fin de ce document).

Compte Google
ou Facebook

Infos personnelles pour utiliser



15

Les exemples de clauses

nous collecterons certaines informations, telles que votre (ou celui de votre enfant autorisé) nom d'utilisateur et les messages envoyés à d'autres utilisateurs. Ces informations ne permettront pas aux autres utilisateurs de vous identifier (ou votre enfant autorisé), à moins que vous (ou votre enfant autorisé) ne choisissez d'utiliser votre (ou celui de votre enfant autorisé) nom réel et autres informations d'identification. Lorsque vous (ou votre enfant autorisé) créez un compte, nous recueillons également d'autres informations (telles que le pays et la langue) qui ne peuvent pas être utilisées pour vous identifier (ou votre enfant autorisé), à moins qu'elles ne soient combinées avec d'autres informations d'identification.

Messages échangés
Langue
Pays

16



Récupération d'autres données

Cor
out
x (e) Les « Cookies » sont de petits fichiers texte qui sont placés sur votre disque dur par un serveur Web lorsque vous (ou votre enfant autorisé) accédez à nos Services. Nous pouvons utiliser les cookies de session et les cookies persistants pour identifier que vous (ou votre enfant autorisé) vous êtes connecté aux Services et pour nous informer de la manière et la période où vous (ou votre enfant autorisé) interagissez avec nos Services. Nous pouvons également utiliser les cookies pour surveiller l'utilisation globale et le routage du trafic web sur nos services et ainsi personnaliser et améliorer nos Services. Les cookies de session sont supprimés lorsque vous (ou votre enfant autorisé) vous déconnectez des Services et fermez le navigateur. Les cookies persistants restent sur votre ordinateur et permettent d'identifier la façon dont vous utilisez les services au fil du temps. Bien que la plupart des navigateurs acceptent automatiquement les cookies, vous pouvez modifier les options de votre navigateur pour cesser d'accepter automatiquement les cookies ou pour vous avertir avant de les accepter. Sachez néanmoins que, suite à ce refus, vous (ou votre enfant autorisé) pourriez ne pas être en mesure d'accéder à toutes les sections ou caractéristiques des Services. Certains prestataires de services tiers que nous engageons (y compris des annonceurs tiers) peuvent également placer leurs propres cookies sur votre disque dur.

17



Partage d'informations

Si des bugs, erreurs, ou d'autres incidents ou problèmes surviennent au cours du fonctionnement ou du développement des Services, alors que vous vous inscrivez pour créer un compte, nous pouvons partager vos données à caractère personnel (ou celles de votre enfant autorisé) avec TPC et / ou TPCI si une telle collaboration s'avère nécessaire pour rechercher, diagnostiquer, corriger et / ou résoudre le problème. Toute information que vous (ou votre enfant autorisé) fournissez directement à TPC et / ou TPCI est soumise à la Politique de confidentialité de l'entreprise applicable. Nous ne sommes pas responsables des politiques et pratiques en matière de confidentialité, de sécurité et / ou contenu de TPC ou TPCI.

Les tiers peuvent accéder et utiliser les données

18



En cas de vente ou fusion

Les informations que nous collectons auprès de nos utilisateurs, y compris les données à caractère personnel, sont considérées comme un actif de l'entreprise. Si nous étions rachetés par un tiers à la suite d'une transaction telle qu'une fusion, une acquisition ou une vente d'entreprise, ou si nos actifs étaient rachetés par un tiers pour cause de faillite ou de cessation de commerce, une partie ou la totalité de nos actifs, y compris vos données à caractère personnel (ou celles de votre enfant autorisé), pourraient être divulguées ou transférées à un tiers acquéreur dans le

 Les données font partie des actifs de l'entreprise

Qui suis je ?

- Niantic



- John Hanke (Google Street View) – Récupération des données Wifi
 - Marius Milner (Hacker accusé) travaille à Niantic sous la direction de John Hanke
 - Financement initial : 20 M\$ via Nintendo et Google
 - A lire les clauses : <https://www.nianticlabs.com/privacy/pokemongo/fr>

20

Une vrai valeur commerciale

- Décembre 2007, (Google, Microsoft, MySpace, AOL et Yahoo!), ont enregistré 336 milliards de données personnelles
 - Yahoo! a récolté 110 milliards de transmissions de données, soit en moyenne 811 (1.700 avec l'ensemble de ses partenaires) informations pour chaque internaute ayant visité un de ses sites durant cette période.
 - 110 milliards de données personnelles en un mois !
 - Dresser un portrait-robot fiable de l'internaute consommateur
 - De 10 à 50 euros !!

21

Tout s'achète

- Site de ventes en ligne sur les clients intéressés par la voyance
- Nom, prénom, adresse, numéro de CB
- 1 euro par personne
- A essayer :



22

Les bases clients protégées ?

- Janvier 2009 : 400 000 fiches du fournisseur d'accès à Internet Orange laissées en libre accès sur Internet via une faille de sécurité
- Octobre 2008 : 30 millions de données de Deutsche Telekom (avec numéros de CB)
- Août 2008 : les données bancaires d'un million de clients en vente sur eBay (pour 44 euros)
- Janvier 2009 : 4 millions de comptes visités par des hackers sur Monster
- Mars 2010 : Fichier SNCF (1 adresse et coordonnées d'un voyageur 8 à 20 euros)



23

De l'anonymisation

- Expérience d'AOL en 2006
- Une liste de 20 millions de recherches d'internautes mis en ligne après avoir été anonymisées
- No. 4417749 a effectué de nombreuses recherches sur « un homme célibataire de 60 ans » et « des informations sur un chiens qui urine partout »
 - En recherchant, localisation (Lilburn, Ga), vue d'un lac, ...
 - Thelma Arnold, a 62-year-old veuve qui vie à Lilburn, Georgie



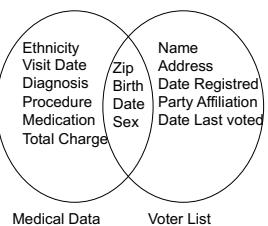
24



De l'anonymisation

- Fichier anonymisé des soins de santé des fonctionnaires de l'état du Massachusetts mis en ligne (L. Sweeney, 1997)
 - La liste électorale de Cambridge, MA (53 805 inscrits)

□ 69 % d'enregistrements uniques par rapport à code postal, date de naissance



Dossier médical du gouverneur du Massachusetts



75

Plan

- Concrètement ?
 - Pourquoi fouiller les données ?
 - Le processus d'extraction
 - Un aperçu de quelques techniques



26

Pourquoi fouiller les données ?

- De nombreuses données sont collectées et entreposées
 - Données du Web, e-commerce
 - Achats dans les supermarchés
 - Transactions de cartes bancaires
 - Les ordinateurs deviennent de moins en moins chers et de plus en plus puissants
 - La pression de la compétition est de plus en plus forte
 - Fournir de meilleurs services, s'adapter aux clients (e.g. dans les CRM)



27

Pourquoi fouiller les données ?

- Les données sont collectées et stockées rapidement (GB/heures)
 - Capteurs : RFID, supervision de procédé
 - Télescopes
 - Puces à ADN générant des expressions de gènes
 - Simulations générant de téraoctets de données



28

Pourquoi fouiller les données ?

- Les techniques traditionnelles ne sont pas adaptées
- Volume de données trop grands (trop de tuples, trop d'attributs)

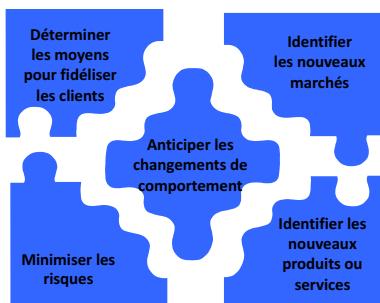
Comment explorer des millions d'enregistrements avec des milliers d'attributs ?
- Besoins de répondre rapidement aux opportunités
- Requêtes traditionnelles (SQL) impossibles

« Rechercher tous les enregistrements indiquant une fraude »
- Croyance dans la présence de données importantes



29

Un enjeu stratégique



30

Qu'est ce que le Data Mining ?

- De nombreuses définitions
 - Processus **non trivial** d'extraction de connaissances d'une base de données pour obtenir de nouvelles données, valides, potentiellement utiles, compréhensibles,
 - Exploration et analyse, par des moyens **automatiques ou semi-automatiques**, de grandes quantités de données en vue d'extraire des motifs intéressants



31

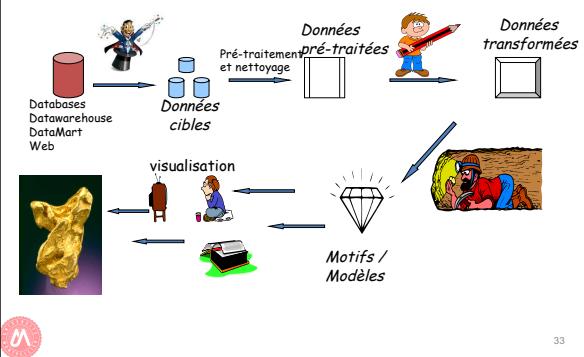
Plan

- Concrètement ?
- Pourquoi fouiller les données ?
- Le processus d'extraction
- Un aperçu de quelques techniques

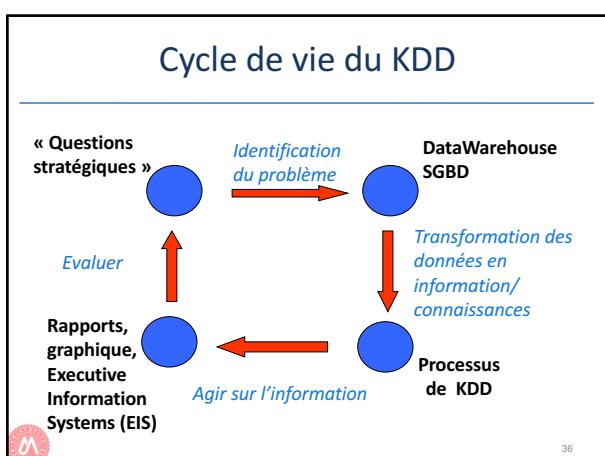
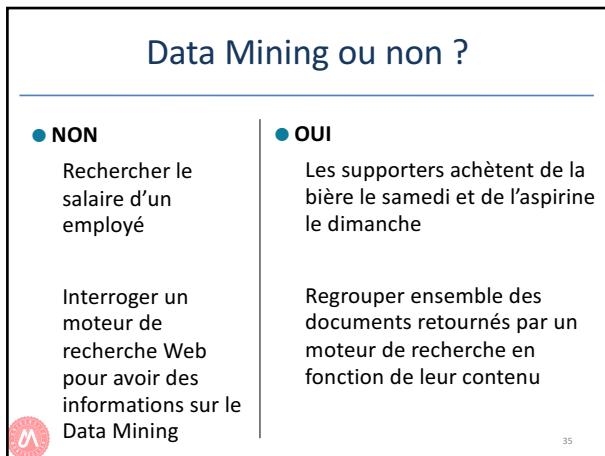
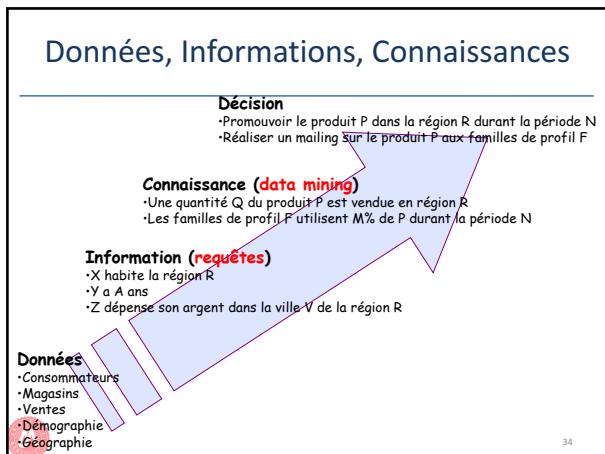


32

Le processus de KDD



33



Applications

- Médecine : bio-médecine, drogue, Sida, séquence génétique, gestion hôpitaux, ...
- Finance, assurance : crédit, prédition du marché, détection de fraudes, ...
- Social : données démographiques, votes, résultats des élections,
- Marketing et ventes : comportement des utilisateurs, prédition des ventes, espionnage industriel, ...
- Militaire : fusion de données .. (secret défense)
- Astrophysique : astronomie, « contact » (:-))
- Informatique : agents, règles actives, IHM, réseau, Data-Warehouse, Data Mart, Internet (moteurs intelligent, profiling, text mining, ...)



37

Quid des données ?

- Grandes Bases de Données ou non ?
- Faut -il échantillonner ?
 - 100 000 enregistrements, 100 Mo par jour
 - 2 Go par jour, 100 Go par heure
 - ... *Déjà les petabyte (2^{30}) ...*
- Différents domaines
 - Bases de Données
 - Intelligence Artificielle (Machine Learning)
 - Statistiques
 - Algorithmique,
 - Visualisation...



38

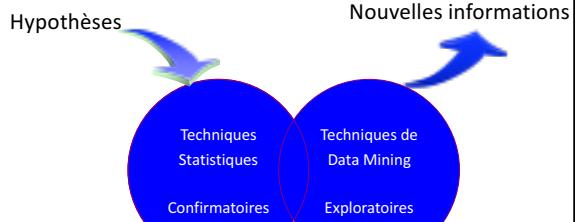
Quid du type de données ?

- Booléennes, Numériques, Symboliques, Multidimensionnelles, Textuelles, Images, ...
- Et ce n'est pas le monde des bizounours
- Gros volumes, Bruitées, Manquantes, Données dynamiques, Données en flots,



39

Data Mining vs Statistiques



40

Attention aux interprétations !



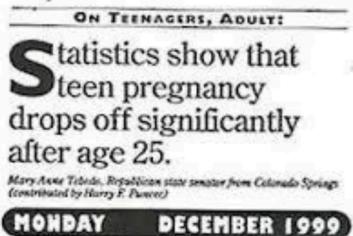
41

Attention aux interprétations !



42

Attention aux interprétations!



Data doesn't create meaning, we do. –Susan Etlinger

43

43

Data Mining vs Machine Learning



Passage à l'échelle ... mais pas seulement

44

Attention aux données!

Quartet d'Ascombe							
I		II		III		IV	
x	y	x	y	x	y	x	y
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91
5,0	5,68	5,0	4,74	5,0	5,73	8,0	6,89

45

45

Attention aux données!

Propriété	Valeur
Moyenne des x	9,0
Variance des x	10,0
Moyenne des y	7,5
Variance des y	3,75
Corrélation entre les x et les y	0,816
Équation de la droite de régression linéaire	$y = 3 + 0,5x$
Somme des carrés des erreurs relativement à la moyenne	110,0

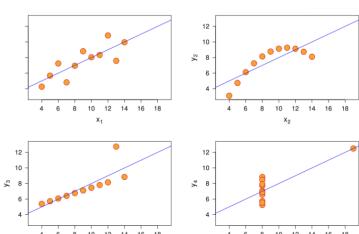


40



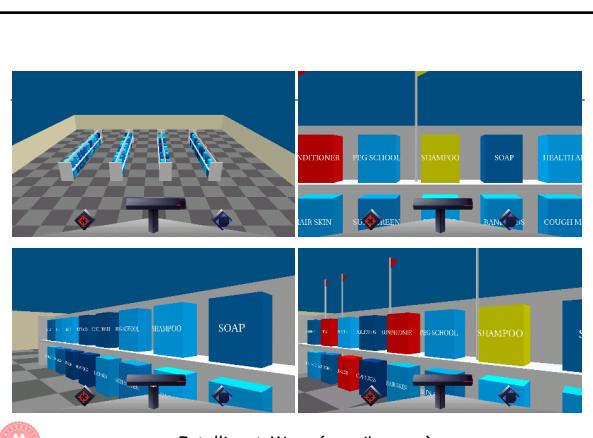
Importance de la visualisation

Quartet d'Ascombe						
I	II	III	IV			
x	y	x	y	x	y	x
10.0	8.04	10.0	9.14	10.0	7.64	8.0
8.0	6.98	8.0	8.14	8.0	6.77	7.0
13.0	7.58	13.0	8.74	13.0	12.74	8.0
9.0	8.81	9.0	8.77	9.0	7.11	8.0
8.0	6.33	10.0	9.26	11.0	7.81	8.0
14.0	9.96	14.0	10.10	14.0	8.84	8.0
6.0	7.24	6.0	6.13	6.0	6.08	8.0
4.0	4.26	4.0	3.10	4.0	5.39	19.0
12.0	10.84	12.0	9.13	12.0	8.15	8.0
7.0	4.82	7.0	7.26	7.0	6.42	8.0
5.0	5.68	5.0	4.74	5.0	5.73	8.0



47

47



Intelligent Miner (www.ibm.com)

48



Plan

- Concrètement ?
- Pourquoi fouiller les données ?
- Le processus d'extraction
- Un aperçu de quelques techniques



49

Les tâches du DM

- Data Mining : de nombreuses tâches possibles ...
 - Classification
créer une fonction qui classe une donnée élémentaire parmi plusieurs classes prédéfinies existantes
 - Régression
créer une fonction qui donne une donnée élémentaire à une variable de prévision avec des données réelles
 - Groupement (clustering)
rechercher à identifier un ensemble fini de catégories ou groupe en vues de décrire les données
 - Résumé
affiner une description compacte d'un sous-ensemble de données
 - Modélisation des dépendances
trouver un modèle qui décrit des dépendances significatives entre les variables
 - Détection de changement et déviation
découvrir les changements les plus significatifs dans les données



50

Les tâches du DM

- Non pas 1 mais n approches ... donc m techniques ...
- 3 approches principales (*R. Agrawal*) vision BD

Classification
Règles d'association
Motifs séquentiels



51

Classification

- division de l'ensemble de données en classes disjointes en utilisant un apprentissage supervisé ou non (clustering)
 - *But* : recherche d'un ensemble de prédicts caractérisant une classe d'objet et qui peut être appliquée à des objets inconnus pour prévoir leur classe d'appartenance.
 - *Exemple* : une banque peut vouloir classer ses clients pour savoir si elle accorde un crédit ou non.
 - *Techniques* : Arbre de décision, réseaux neuronaux, ...



52

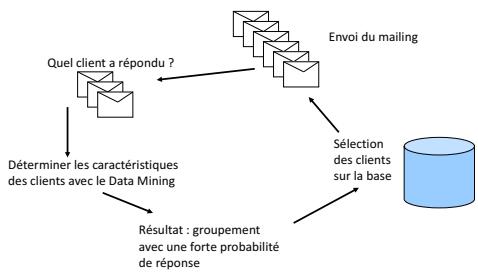
Le mailing

- Classification... un exemple d 'utilisation
 - un cadeau est envoyé par mailing. Un envoi sans réponse coûte 50 € et une réponse assure 100 €.
 - Pas d 'envoi de mailing à un client qui aurait répondu : perte de 100 €.



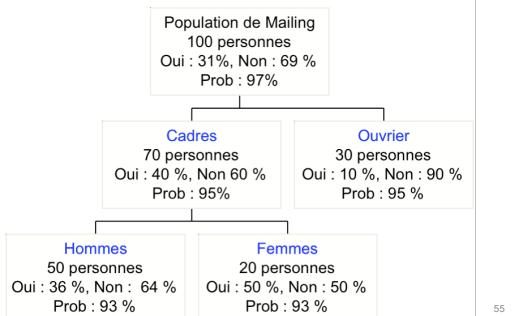
53

Le mailing

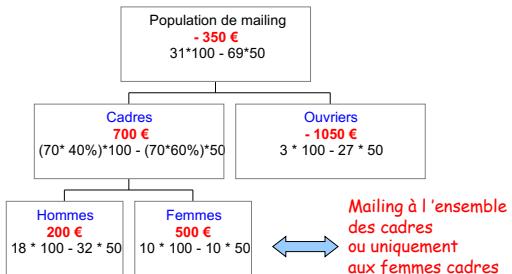


54

Résultat du mailing



Quantification



Evaluation

Matrice de coûts

		OBSERVE			TOTAL
Prédit	Payé	Payé	Retardé	Impayé	
Payé	80	15	5	100	
Retardé	1	17	2	20	
Impayé	5	2	23	30	
TOTAL	86	34	30	150	

Validité du modèle : nombre de cas exacts
(=somme de la diagonale) divisé par le nombre total :
 $120/150 = 0.8$

57

Recherche de motifs fréquents

- Qu'est ce qu'un motif fréquent ?
 - Un motif (ensemble d'items, séquences, arbres, ...) qui interviennent fréquemment ensemble dans une base de données [AIS93]
 - Les motifs fréquents : une forme importante de régularité
 - Quels produits sont souvent achetés ensemble ?
 - Quelles sont les conséquences d'un ouragan ?
 - Quel est le prochain achat après un PC ?



58

Recherche de motifs fréquents

- Analyse des associations
 - Panier de la ménagère, cross marketing, conception de catalogue, analyse de textes
 - Corrélation ou analyse de causalité
 - Clustering et Classification
 - Classification basée sur les associations
 - Analyse de séquences
 - Web Mining, détection de tendances, analyses ADN
 - Périodicité partielle, associations temporelles/cycliques



59

- Des questions ?



60
