1

Extraction de Connaissances
Fouille de Données

http://www.lirmm.fr/~poncelet/FMIN311.html

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

Des choses entendues

- A quoi l'extraction de connaissances, la fouille de données, l'aide à la décision peuvent ils bien servir ?
  - « oui d'accord on stocke des données et on voudrait avoir des connaissances mais bon … rien de très concret »
  - « c'est super grâce à cela je suis capable d'une part de comprendre mes données mais surtout de savoir sur quelles variables je dois agir »
- Oups … contradictoire

2

Des choses entendues

- Quelques choses d'abstrait ? Concret ?
  - « franchement je ne vois pas la différence par rapport à ce que je faisais avant. Une requête et je sais »
  - « c'est complètement différent de ce que nous avions avant. Grâce à ces techniques je suis capable de réellement découvrir des informations utiles … au fait on dit informations ou connaissances »
- Oups … contradictoire

3

Des choses entendues

- Un domaine uniquement académique ? Un domaine uniquement industriel ?
  - « de toute façon c'est que pour les industriels, c'est uniquement pour gagner de l'argent, il n'y a rien de scientifique ou de réellement académique là dedans »
  - « L'un des articles de recherche les plus cités au monde est un article de fouille de données »
- Oups …contradictoire

4

Théorique ?

• Théorème : Soit s une séquence et CS un ensemble de contraintes de fréquences pour toutes les séquences s' telles que $s' \leq s$. S'il existe une base de données transactionnelles D satisfaisant cet ensemble de contraintes alors la borne inférieure pour la fréquence de la séquence s, notée $LB(S)$, est égale à 0.

• Corollaire : Toutes les représentations condensées basées sur le support issues des k-libres (non dérivables, 0-libre, disjonctifs libres, …) sont inintéressantes dans le cadre des motifs séquentiels

Best Papers - PKDD 08

- Conséquences : au niveau de la communauté internationale, il est prouvé théoriquement que ce n'est pas possible de trouver mieux :)

5

Théorique ?

Théorème : Soit s une séquence et S_D la taille de l'échantillon, alors nous avons : $Pr[e(s,S_D) > \varepsilon] \leq \delta$ si $|S_D| \geq \ln(2/\delta)1/(2\varepsilon^2)$ où $e(s,S_D)$ correspond au taux d'erreur absolu ($e(s,S_D) = |Support(s,S_D) - Support(s,D)|$)

Généralisation de l'approche Sampling de Toivonen [96] pour les itemsets

IEEE International Conference On Data Mining (ICDM 07)

ε	δ	$\lvert S_D \rvert$
0.01	0.01	26492
0.01	0.001	38005
0.001	0.01	2649160
0.001	0.0025	3333333

6

Des choses entendues

- Une mode ?
  - « en fait on fait cela depuis longtemps ce n'est qu'une mode boostée par les industriels. »
  - « Le MIT considère que la fouille de données au sens large est l'un des grands challenges des 10/20 prochaines années. BIG DATA = gérer et extraire de la connaissance de gros volumes de donnés en flux»
- Oups …contradictoire

7

Des choses entendues

- Du réchauffé ?
  - « extraire de la connaissance, cela fait des années qu'on le fait. Franchement il n'y a rien de nouveaux dans ce qui se fait actuellement »
  - « Les approches existantes n'ont jamais été capables de gérer la grande quantité de données disponibles. Elles n'ont pas de présupposé. Les nouvelles avancées en fouille de données permettent enfin de pouvoir avoir les mécanismes pour le faire. Enfin nous pouvons aborder le passage à l'échelle. Tout ne tient pas en mémoire centrale»
- Oups …contradictoire

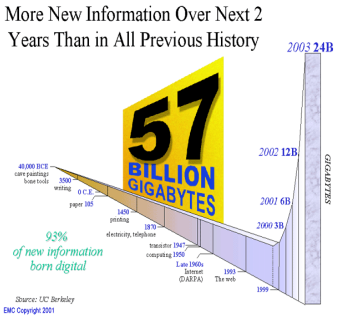
8

Alors ?

- A vous de vous faire votre opinion ☺
- En tout cas on s'amuse bien ☺

9

Travailler sur de gros volumes



- (une petite paranthèse …..)

10

Quelques chiffres

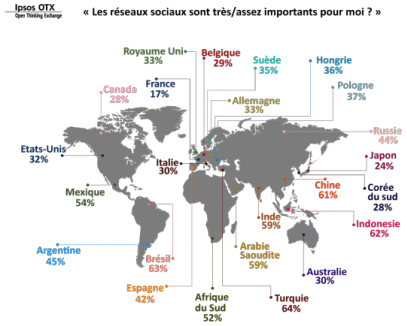
- 2,3 milliards dans le monde
- 41,2 millions en France
- 32 millions sont inscrits sur au moins un réseau social
- 52 % ont entre 25 et 45 ans

- Facebook : 26 000 000 utilisateurs Français sur 1,6 milliards de membres !!
- En moyenne les utilisateurs ont 177 amis :)

(Sources Factory.com (2013))

11

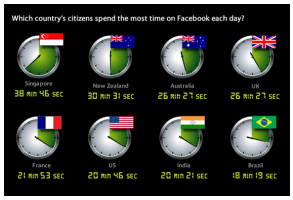
De l'utilisation des sites de réseaux sociaux



12

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. »

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

Les amis de mes amis

- Entretien avec Alex Türk, président de la Cnil (Commission nationale de l'informatique et des liberté).
- **« 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 arrêtez de vous flatter, cela ne fait que 5 mois que vous travaillez ici, n'avez pas remarqué que je suis gay. Ensuite le travail de m… comme vous dites est le travail pour lequel je vous paye […]. Vous semblez avoir oublié qu'il vous restait encore deux semaines de travail en période d'essai. **Ne prenez pas la peine de revenir demain**. »
- Son patron était en relation sur Facebook

Source Grande Bretagne - Août 2009

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 ….

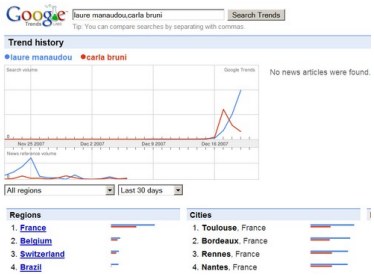
Une expérience

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

17

Les moteurs de recherche

- Les photos de Laure Manaudou - décembre 2007



18

Difficile ?

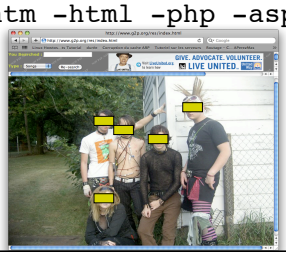
19

Non - google requêtes complexes

La requête google :

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

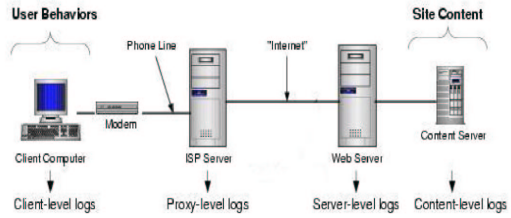
XXXXMyBestFriendsXXXXXX.jpg



20

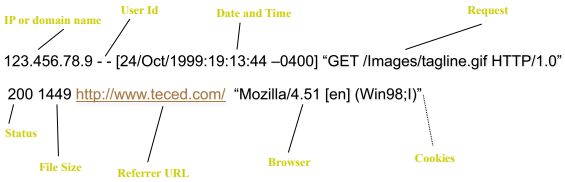
Log ou Logs ?

Information sur les chemins de navigation dans les fichiers logs



21

Web logs + ?



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

22

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 !!

23

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 :)

24

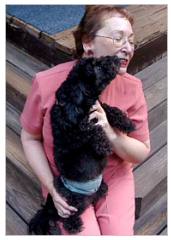
Les bases clients protégées ?

- 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 Deutche 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)

25

De l'anonymisation

- Expérience d'AOL en 2006
- Une liste de 20 millions de recherche 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



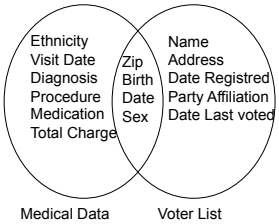
26

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 Cambrige, MA (53 805 inscrits)

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

Dossier médical du gouverneur du Massachusetts



27

Plan

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

28

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)

29

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

30

Pourquoi fouiller les données ?

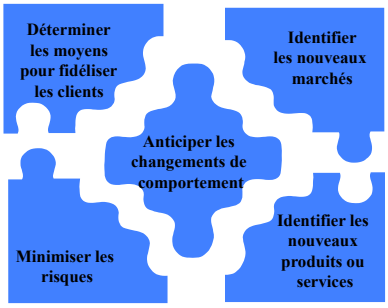
- Les techniques traditionnelles ne sont pas adaptées
- Volume de données trop grands (trop de tuples, trop d'attributs)

  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

31

Un enjeu stratégique



32

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

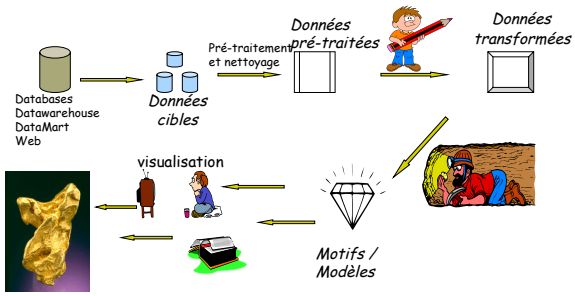
33

Plan

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

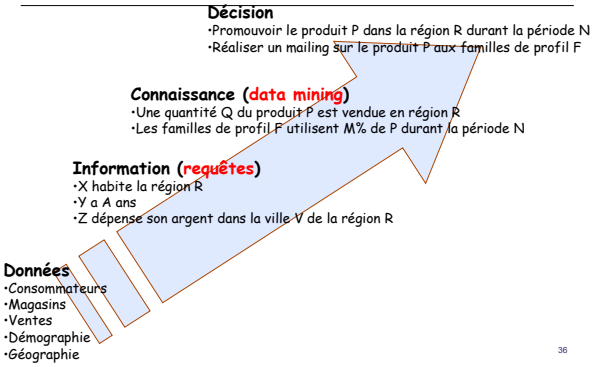
34

Le processus de KDD



35

Données, Informations, Connaissances



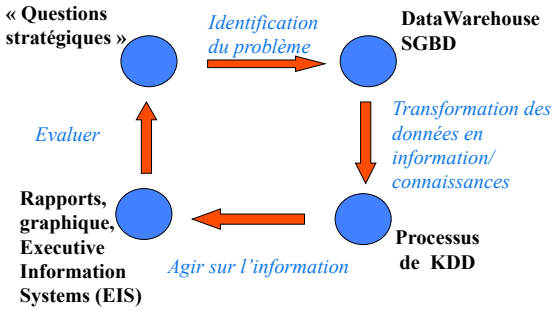
36

Data Mining ou non ?

- **NON**
  - Rechercher le salaire d'un employé
  - Interroger un moteur de recherche Web pour avoir des informations sur le Data Mining
- **OUI**
  - Les supporters achètent de la bière le samedi et de l'aspirine le dimanche
  - Regrouper ensemble des documents retournés par un moteur de recherche en fonction de leur contenu

37

Cycle de vie du KDD



38

Applications

- Médecine : bio-médecine, drogue, Sida, séquence génétique, gestion hôpitaux, ...
- Finance, assurance : crédit, prédiction du marché, détection de fraudes, …
- Social : données démographiques, votes, résultats des élections,
- Marketing et ventes : comportement des utilisateurs, prédiction 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, …)

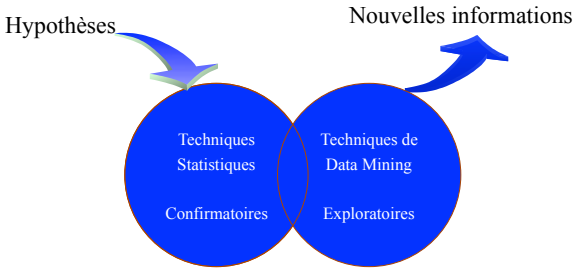
39

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^{50}) ...*
- Différents domaines
  - Bases de Données
  - Intelligence Artificielle (Machine Learning)
  - Statistiques
  - Algorithmique,
  - Visualisation…

40

Data Mining vs Statistiques

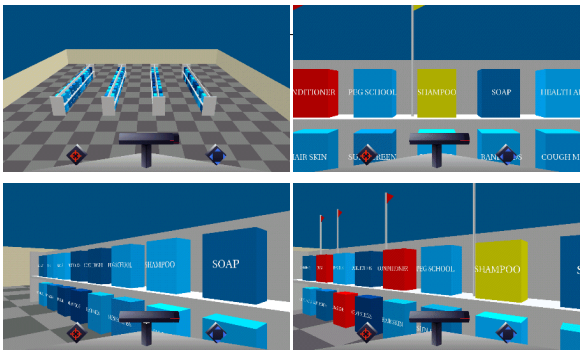


41

Data Mining vs Machine Learning

Passage à l'échelle

42



Intelligent Miner (www.ibm.com)

43

Les règles d'association

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, ….

44

Plan

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

45

Les tâches du DM

- Data Mining : de nombreuses tâches possibles …
  - Classification
    créer une fonction qui classifie 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

46

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

47

Classification

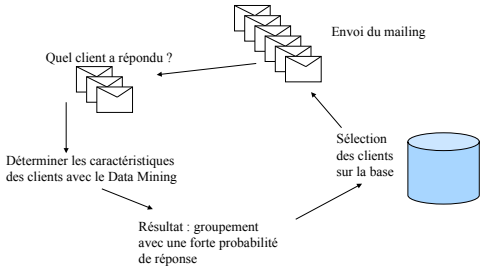
- division de l'ensemble de données en classes disjointes en utilisant un apprentissage supervisé ou non (clustering)
  - *But* : recherche d'un ensemble de prédicats caractérisant une classe d'objet et qui peut être appliqué à des objets inconnus pour prévoir leur classe d'appartenance.
  - *Exemple*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

48

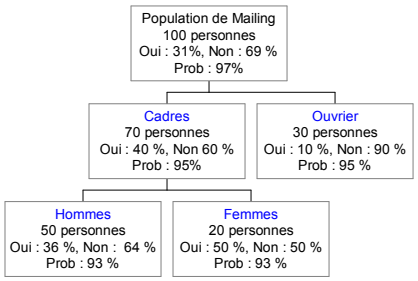
Le mailing

- 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 €.

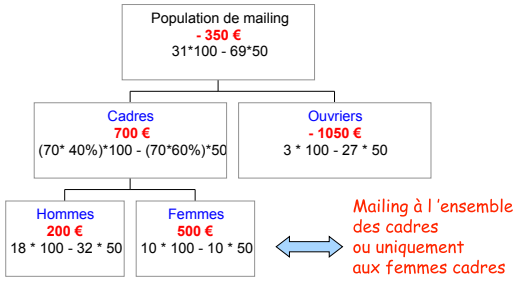
Le mailing



Résultat du mailing



Quantification



Evaluation

Matrice de coûts

Prédit ↓	OBSERVE Payé	Retardé	Impayé	TOTAL
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

53

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?

54

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

55