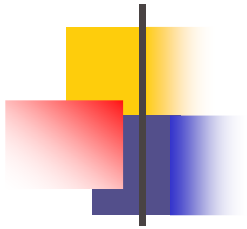
Découverte de connaissances dans les cubes de données

Rokia Missaoui, Ph.D.
http://w3.uqo.ca/missaoui

Laboratoire LARIM

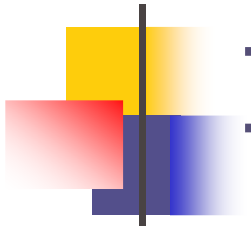
Université du Québec en Outaouais (UQO)

Mai 2005



Plan

- Introduction
- Fouille de données *: Data mining* (DM)
- Entreposage de données *: Data warehousing*
- Exploration des cubes de données
- Modèles log-linéaires
- Nos travaux
- Survol des travaux sur le DM dans les cubes



Introduction

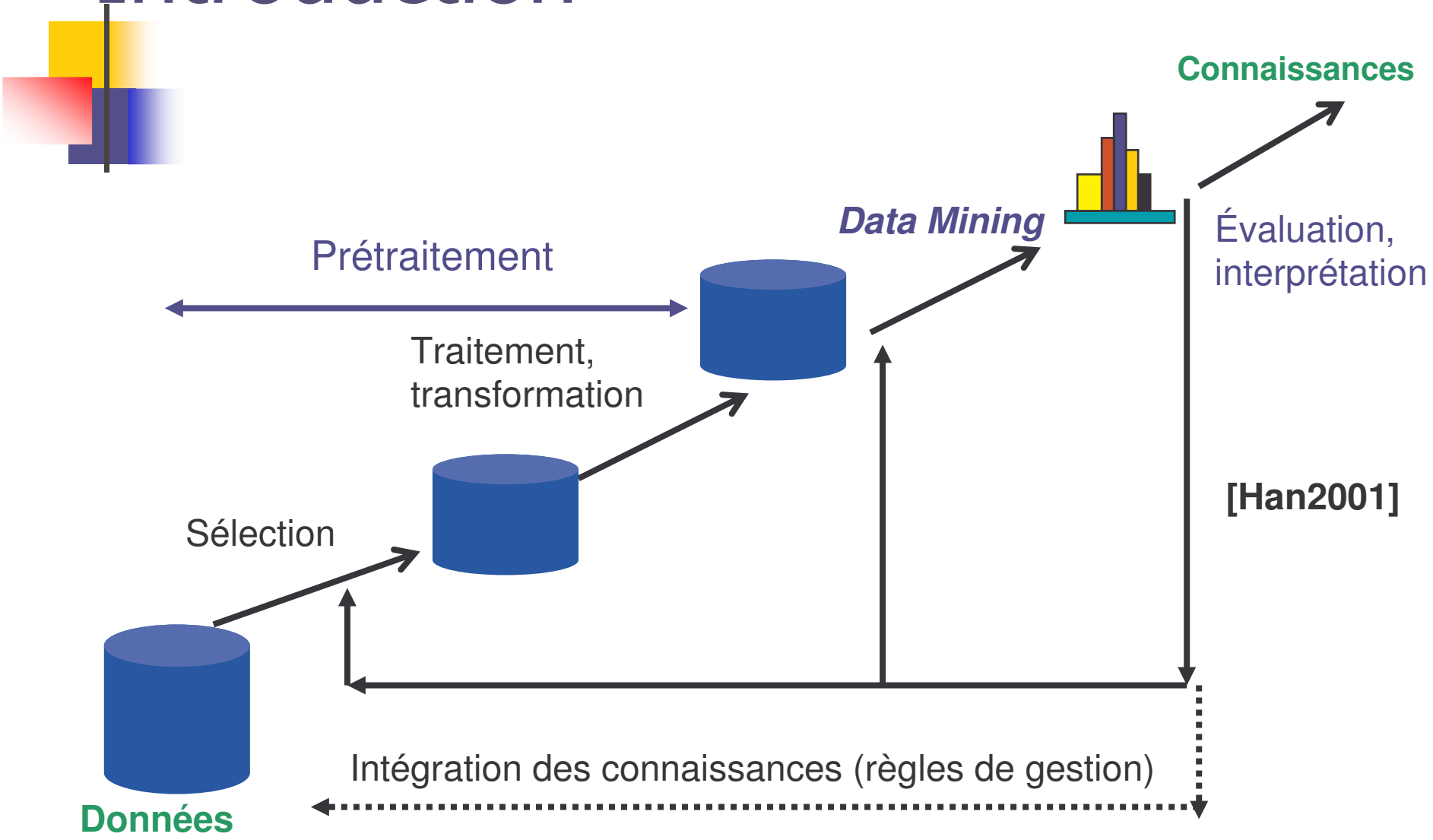
- *We are drowning in data, but starving for knowledge!* (Han 2001)

- Solutions
  - Fouille de données
  - Entreposage de données
  - Intégration des deux technologies

Introduction

- Découverte de connaissances dans le bases de données (KDD)
  - Processus non trivial d'identification, à partir des données, de *patterns*
    - valides (pour de nouvelles données avec un bon degré de certitude), nouveaux
    - potentiellement utiles (i.e, devraient conduire vers des décisions utiles)
    - ultimement compréhensibles (par des humains)
  - Fouille de données : étape du processus KDD

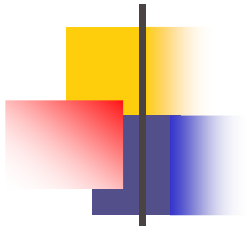
Introduction



Découverte de connaissances dans les bases ou cubes de données

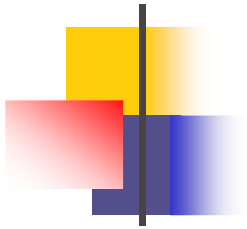
Problématique

- Découverte de connaissances dans les cubes de données (DCCD)
  - Problème au niveau des données
    - Grand volume
    - Forte dimensionnalité
    - Multidimensionnalité
  - Problème au niveau des traitements
    - Les techniques OLAP permettent un traitement analytique des données mais ne permettent pas de déceler des valeurs aberrantes (*outliers*), des associations, des groupements, …



Fouille de données

- Trois principales formes
  - Prédiction
    - Apprentissage supervisé
    - Prédiction de la valeur d'une variable (régression) ou de la classe d'appartenance (classification)
    - Réseaux bayésiens, arbres de décision, réseaux de neurones
    - Ex. identification des fraudeurs et des clients à haut risque
  - Découverte
    - Apprentissage non supervisé, analyse exploratoire

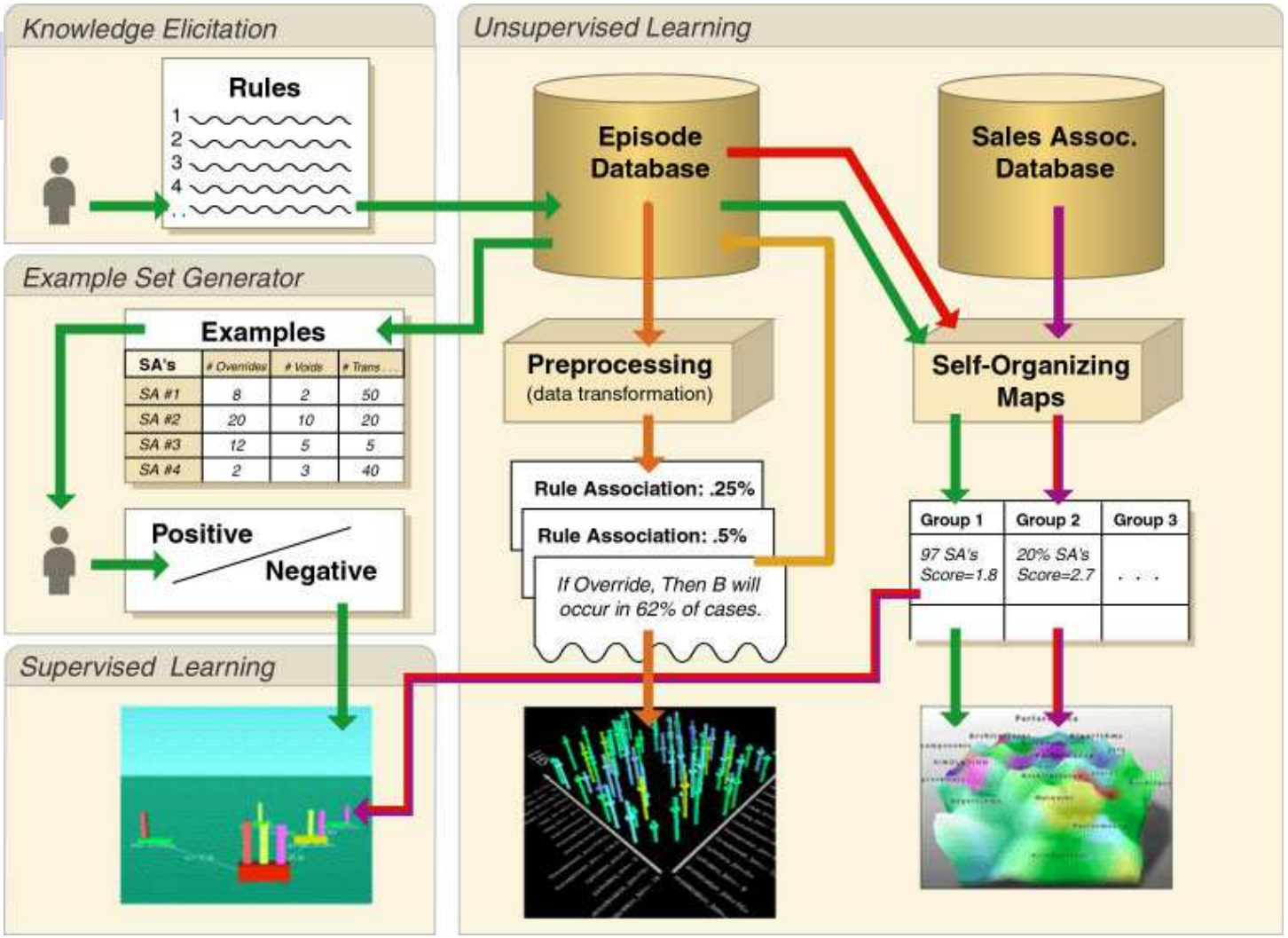


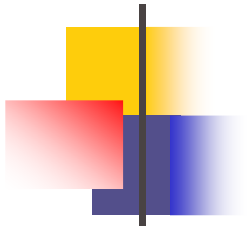
Fouille de données

- Trois principales formes (suite)
  - Découverte
    - Règles d'association
      - Ex. Achat du lait $\Rightarrow$ Achat du beurre [sup. =.5, conf.= .75]
    - Patrons séquentiels
      - Ex. Si le client acquiert un magnétoscope, alors il a une probabilité de 75% d'acheter un caméscope dans un délai de 4 mois
    - Analyse de grappes (*cluster Analysis*)
      - Ex. Segmentation de la clientèle par genre et âge
  - Détection de déviation
    - Valeurs exceptionnelles (*outliers*), analyse de tendances

Fouille de données

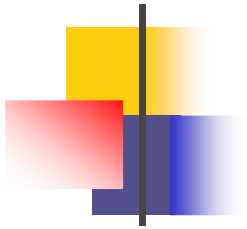
(Welge 2003)





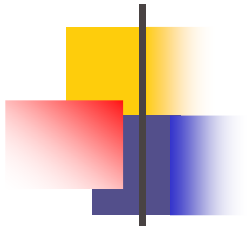
Fouille de données

- Exigences envers les méthodes de DM
  - Performance
  - Passage à l'échelle (*Scalability*)
  - Mise à jour incrémentale des résultats du DM
  - DM "à la demande" ou guidé par des contraintes
- DM dans les applications
  - Développement de solutions de DM taillées sur mesure pour les besoins de l'organisation
  - DM invisible : fonction DM encastrée (*built-in*) dans les applications (Ex. Amazon.com)



Entrepôts de données

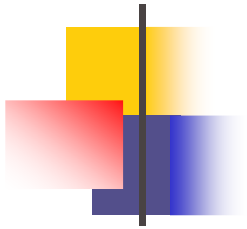
- Entrepôts de données - *Data warehouses (DW)*
- Objectifs
  - Offrir un accès à une version agrégée et historisée de l'ensemble des données de l'entreprise
  - Offrir des outils d'aide à la décision (OLAP)
- Applications possibles
  - Marketing, analyse financière, gestion de la relation client (*CRM*), analyse de profitabilité, analyse de la qualité, gestion des accès au Web, gestion médicale, etc.
- Traitements possibles
  - *OLAP*
  - *Data mining*



Entrepôts de données

- Entrepôts de données (*Data warehouse*)
  - BD orientée sujet, intégrée et non volatile de données temporelles
  - Système informationnel contenant surtout des données agrégées

- Magasins de données (*Data mart*)
  - Collection de données logiquement apparentées répondant aux besoins d'un groupe spécifique d'usagers ou d'une unité administrative
  - Souvent un sous-ensemble de l'entrepôt



Entrepôts de données

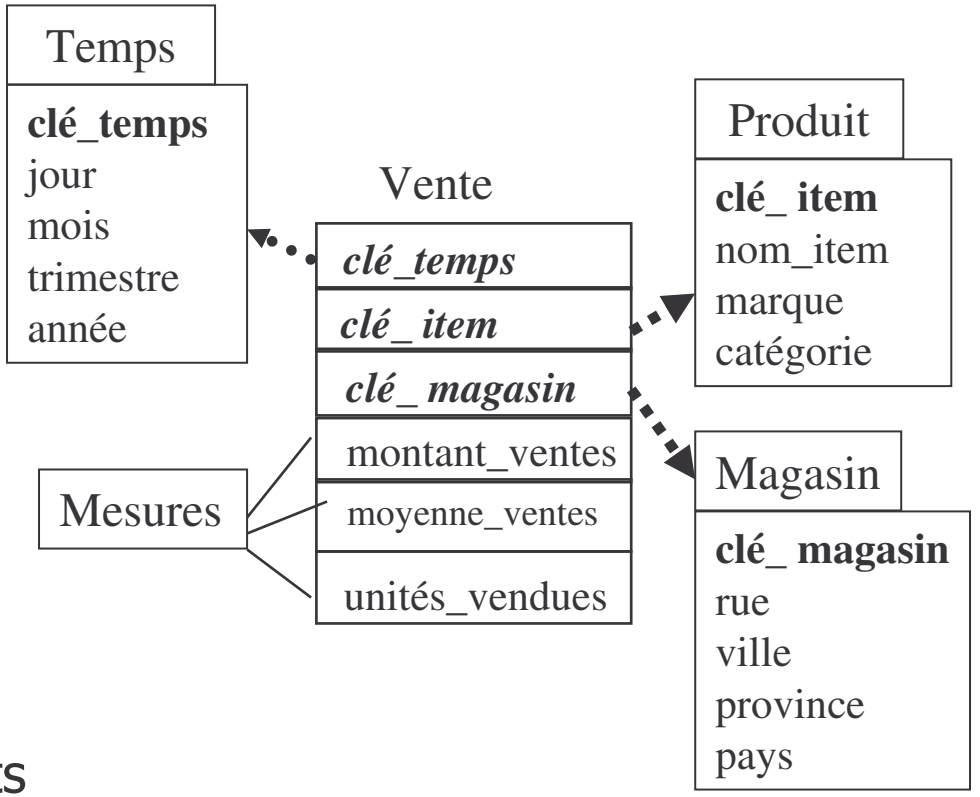
- Trois types d'applications
  - Traitement de l'information
    - Requêtes, analyses statistiques de base, rapports, diagrammes et graphiques
  - Traitement analytique
    - Analyse multidimensionnelle, opérations OLAP
  - *Data mining*
    - Découverte de connaissances (patrons cachés)
    - Associations, modèles analytiques, classification et prédiction, et mécanismes de visualisation

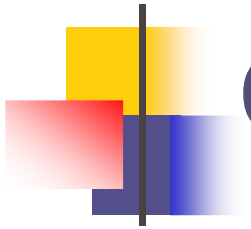
Entrepôts de données

Types de modèles

- Étoile (*Star*)
  - Table centrale (*fact table*) normalisée
    - Clé primaire composée
    - Mesures (indicateurs)
  - Tables de dimensions non normalisées
- Flocons de neige
  - Tables de dimensions non normalisées
- Galaxie - constellation de faits
  - Plusieurs tables de dimension

Exemple : Schéma en étoile





Cubes de données

- Point de vue statistique
  - Un cube de données est une représentation multidimensionnelle d'un tableau multivarié dans un entrepôt de données
- Utilité
  - Moyen efficace de représenter simultanément deux ou plusieurs caractères observés sur une même population
- Composition
  - Dimension (variable de catégorie). Ex. produit
  - Modalités (membres) de la dimension. Ex. lait, fromage,
  - Mesure (ex : montant des ventes, unités vendues, …)
- Hiérarchie de dimensions
  - Exemple : Item -> Ligne -> Groupe de produits

Cubes de données

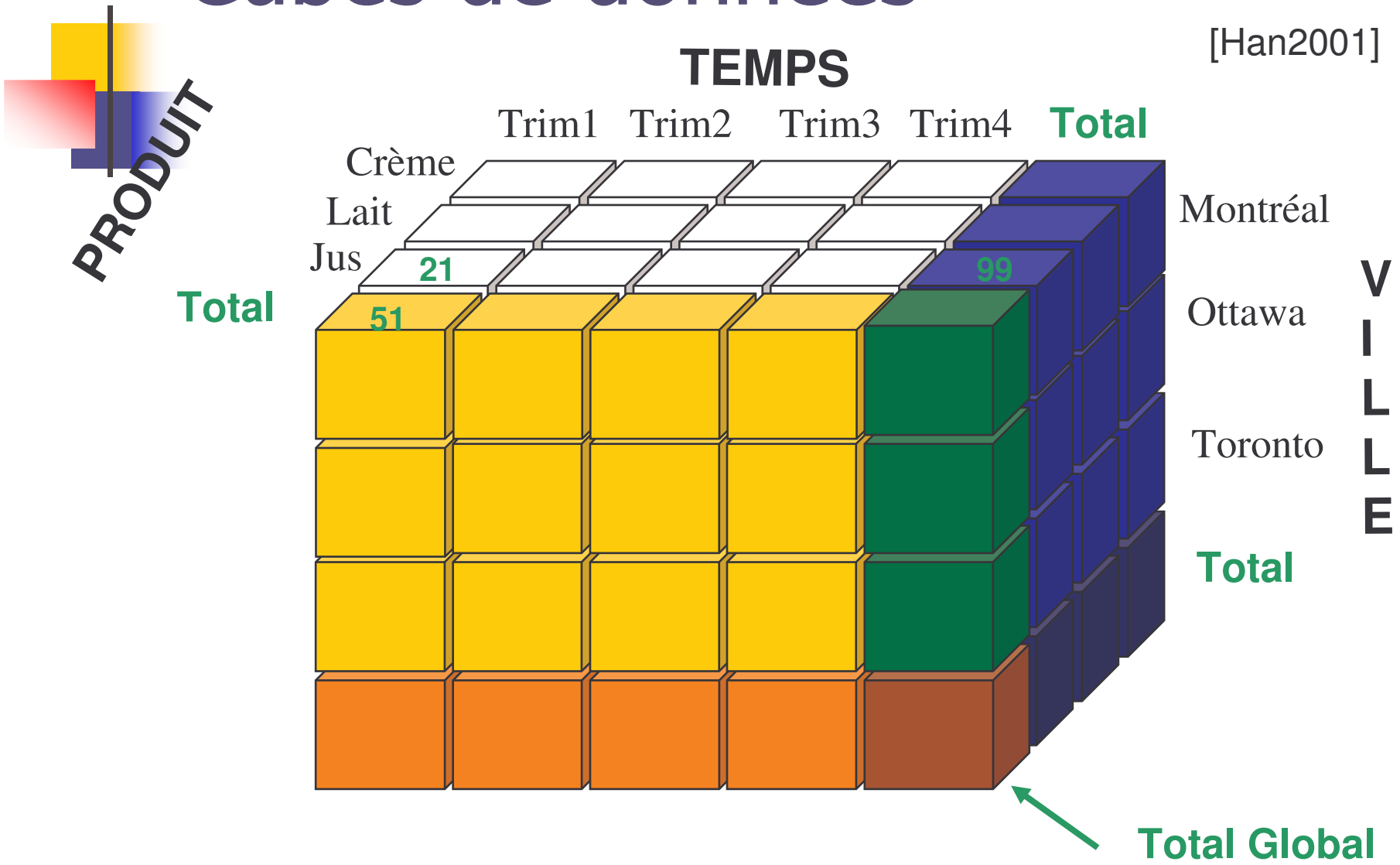
M : Montréal, O : Ottawa, T : Toronto

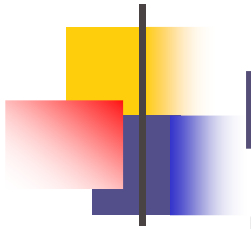
Temps / Ville / Produit	Trim1			Trim2			Trim3			Trim4			**Total**			
	M	O	T	M	O	T	M	O	T	M	O	T	M	O	T	Tot
Crème	**8**			**4**			**6**			**9**			**27**			
Lait	**22**			**23**			**19**			**29**			**93**			
Jus	**21**			**24**			**25**			**29**			**99**			
Total	**51**			**..**												

Répartition des ventes par produit, temps et ville

Cubes de données

[Han2001]





Exploration des cubes de données

- Exploration guidée par les hypothèses
  - Exploration par l'usager à l'aide des opérations OLAP (*drill-down, roll-up, slice, dice, pivot*, **..**)
- Avantages
  - Permet de visualiser les données selon diverses perspectives
- Inconvénients
  - Espace de recherche trop grand
    - Pour un cube de **n** dimensions et $\mathbf{L_i}$ niveaux de hiérarchie pour la dimension $\mathbf{D_i}$

$$T = \prod_{i=1}^{n} (L_i + 1)$$

    - **Exemple :** un cube de 8 dimensions avec des hiérarchies de dimension de 7 niveaux offre 1,6 millions (8^8) cuboïdes possibles

Exploration des cubes de données

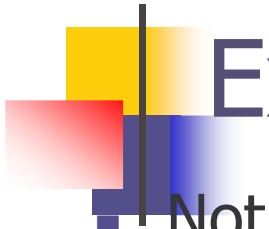
- Exploration guidée par la découverte
  - Travaux de Sarawagi *et al.*
  - Modèles log-linéaires et optimisation du calcul des coefficients
  - Découverte d'exceptions (écarts significatifs entre la valeur observée et la valeur estimée)
  - Représentation visuelle (avec contraste de couleurs) des exceptions

Exploration des cubes de données

- Extension aux règles d'association
  - Travaux de Nestorov *et al.*
  - Génération des règles d'association dans les cubes
  - Exemples
    - Achat d'un sac de couchage → achat d'une tente dans la région nord et pour la saison d'été
    - Achat de couches pour bébés → achat de bière dans l'état de la Californie pour les clients de sexe masculin particulièrement les jeudis soir

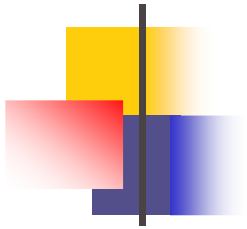
Exploration des cubes de données

- Opérateur RELAX : travaux de Sathe *et al.*
- Objectifs
  - Vérifier si une tendance est vraie dans la plupart des cas et à divers niveaux de granularité
  - Identifier les cas exceptionnels (i.e., ne vérifient pas la tendance)
- Exemple
  - Hausse des ventes de portables Powerbook G4 entre 2003 et 2004 pour la ville de Montréal
  - Vérifier si la tendance se maintient au Québec puis au Canada pour ce type d'ordinateur, puis pour les ordinateurs Apple



Exploration des cubes de données

- Notion de *cubegrade*
  - Travaux de Imielinski *et al.*
  - *Cubegrade* : généralisation des règles d'association
  - Analyse des changements de mesures à la suite d'une modification au niveau des valeurs des dimensions du cube.
  - Utile pour « *What if* »
    - Ex. Si le client achète non seulement du lait mais également des céréales, alors sa consommation moyenne de lait augmente de 20% quand son âge est inférieur à 40 ans et baisse de 5% quand son âge est supérieur ou égal à 40 ans.



Modélisation log-linéaire

- Modélisation des tableaux de fréquences
  - Deux ou plusieurs variables qualitatives
  - Analyse du lien entre deux ou plusieurs variables en prenant le logarithme des fréquences des cellules du tableau
- Complément de l'analyse des correspondances
- Puissant outil pour
  - Tester les liaisons et interactions entre des variables qualitatives
  - Identifier les valeurs extrêmes
- À l'opposé de la régression logistique
  - Aucune distinction n'est faite entre les variables explicatives et les variables expliquées.

Modélisation log-linéaire

- Repose sur l'approximation des fréquences observées des cellules par des fréquences estimées

- Théoriquement, pour une valeur $f_{i1 i2 \ldots in}$ dans un cube C à la position i_r de la $r^{ième}$ dimension d_r ($1 \le r \le n$), il est possible de calculer la valeur estimée $F_{i1 i2 \ldots in}$ comme une fonction *s* de contribution de plusieurs niveaux d'agrégation par:

$$F_{i1 i2 \ldots in} = s\left(\gamma^{G}_{(ir \mid d_r \in G)} \mid G \subseteq \{d_1, d_2, \ldots d_n\}\right)$$

- Exemple : cas de 3 dimensions A, B, C

$$(F_{ijk}) = s\left(\gamma, \gamma^{A}_{i}, \gamma^{B}_{j}, \gamma^{C}_{k}, \gamma^{AB}_{ij}, \gamma^{AC}_{ik}, \gamma^{BC}_{jk}, \gamma^{ABC}_{ijk}\right)$$

- Forme additive

$$\log(F_{ijk}) = \lambda + \lambda^{A}_{i} + \lambda^{B}_{j} + \lambda^{C}_{k} + \lambda^{AB}_{ij} + \lambda^{AC}_{ik} + \lambda^{BC}_{jk} + \lambda^{ABC}_{ijk}$$

Modélisation log-linéaire

Propriété de la forme additive :

- La somme des paramètres λ vaut zéro au niveau des variables

Exemple : Pour un modèle à trois dimensions A, B et C

$$\lambda_{.}^{A} = \lambda_{.}^{B} = \lambda_{.}^{C} = 0$$

$$\lambda_{i.}^{AB} = \lambda_{.j}^{AB} = \lambda_{i.}^{AC} = \lambda_{.k}^{AC} = \lambda_{j.}^{BC} = \lambda_{.k}^{BC} = 0$$

$$\lambda_{ij.}^{ABC} = \lambda_{i.k}^{ABC} = \lambda_{.jk}^{ABC} = 0$$

Nombre de paramètres :

Pour un modèle à trois dimensions (variables) : A, B et C
où A possède n modalités, B possède m modalités et C possède p modalités, on a :

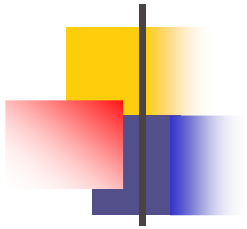
1 + (n-1) + (m-1) + (p-1) + (n-1)(m-1) + (n-1)(p-1) + (m-1)(p-1) + (n-1)(m-1)(p-1) = m*n*p

Formes du modèle log-linéaire

- Forme hiérarchique
  - Si le modèle contient un effet de haut niveau, il va également contenir tous les autres effets de bas niveau
- Forme non hiérarchique
  - Les effets de bas niveaux ne sont pas nécessairement présents
  - Exemple

$$\log(\ F_{ijk}\ ) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ijk}^{ABC}$$

- L'utilisation d'une approche non hiérarchique dépend fortement de la nature des données et de l'application
- De façon générale, il est préférable d'utiliser une approche hiérarchique.



Formes du modèle log-linéaire

- Forme hiérarchique
  - Si le modèle contient un effet de haut niveau, il va également contenir tous les autres effets de bas niveau
- Forme non hiérarchique
  - Les effets de bas niveaux ne sont pas nécessairement présents
  - Exemple: Trois items A, B et C sont vendus séparément. Si le troisième item est gratuit lorsque le client en achète deux, alors il existe une interaction triple entre A, B et C mais aucune interaction double entre les produits

$$\log(\ F_{ijk}\ ) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ijk}^{ABC}$$

- L'utilisation d'une approche non hiérarchique dépend fortement de la nature des données et de l'application
- De façon générale, il est préférable d'utiliser une approche hiérarchique.

Modèle saturé versus non saturé

- **Modèle saturé**
  - Tous les effets possibles des paramètres sont représentés dans le modèle (il y a autant de coefficients qu'il y a de cellules dans le cube)
  - Modélise parfaitement les fréquences observées puisque tous les effets sont représentés.

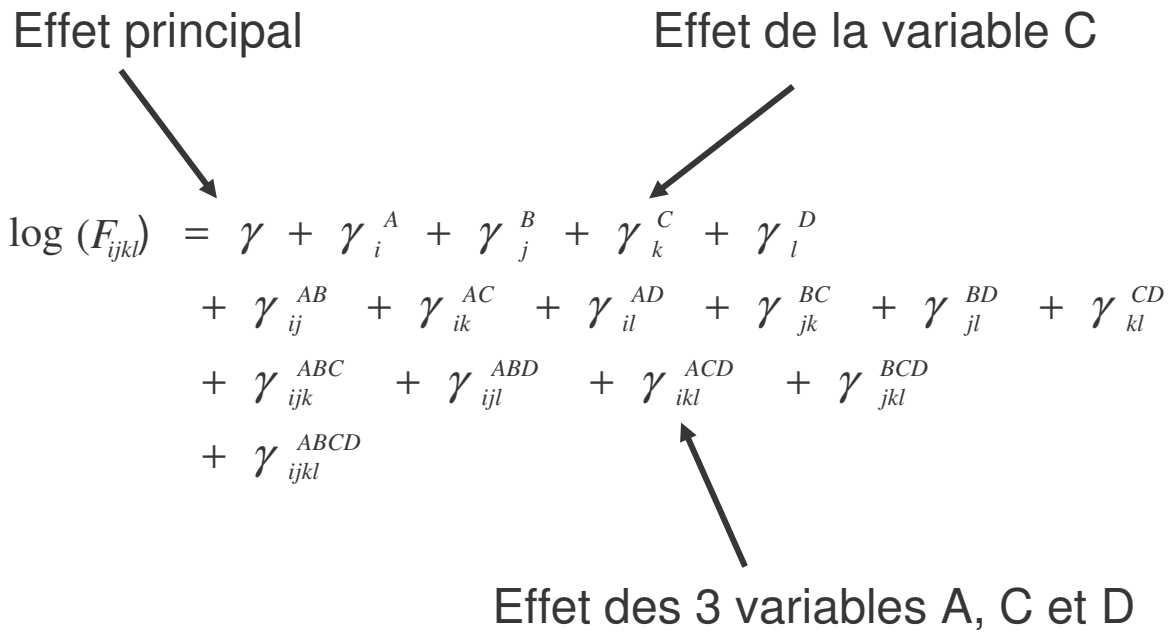
- **Modèle non saturé**
  - Au sein des modèles non saturés, les effets non significatifs des paramètres ne sont pas représentés. Exemple :

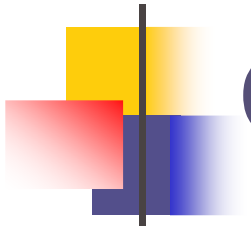
$$\log(F_{ijk}) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C$$

- **Modèle parcimonieux**
  - Modèle ayant le moins d'effets possibles (coefficients) sans une perte excessive de précision

Modélisation log-linéaire

- **Exemple de modèle saturé**





Calcul des coefficents

- Deux principales méthodes
  - Méthode d'ajustement proportionnelle et itérative (*iterative proportional fitting method*)
  - Calcul des coefficients directement des valeurs présentes dans les cellules (Sarawagi et al.)
    - La méthode suppose que les logarithmes des valeurs suivent une loi normale de même variance.
    - Cette méthode n'est pas très efficace en présence de nombreuses valeurs extrêmes.

Calcul des coefficients - Exemple

- Méthode de Sarawagi et al.

P	M	A	
		2000	2001
Café	Provigo	7	6
	Loblaws	8	6
Thé	Provigo	3	5
	Loblaws	9	7

1. **Calcul de la moyenne générale :**

$$\lambda = l_{...} = 6,375$$

2. **Calcul des interactions simples :**

$$\lambda_1^P = l_{i..} - \lambda = {}^{(7+6+8+6)}\!/_{4} - 6,375 = 0,375$$

$$\lambda_2^P = -0,375 \quad \lambda_1^M = -1,125 \quad \lambda_2^M = 1,125$$

$$\lambda_1^A = 0,375 \quad \lambda_2^A = -0,375$$

3. **Calcul des interactions doubles :**

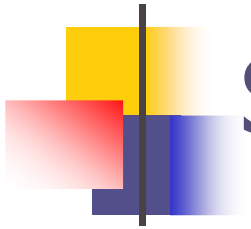
$$\lambda_{11}^{PM} = l_{ij.} - \lambda_1^P - \lambda_1^M - \lambda = 0,875$$

…

4. **Calcul des interactions triples :**

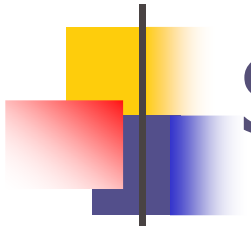
$$\lambda_{111}^{PMA} = l_{ijk} - \lambda_1^P - \lambda_1^M - \lambda_1^A - \lambda_{11}^{PM} - \lambda_{11}^{PA} - \lambda_{11}^{MA} - \lambda = 0,375$$

…



Sélection du modèle

- Deux objectifs conflictuels
  - Parcimonie
    - Représentation des données par un nombre réduit de paramètres
  - Fidélité de la description
    - Recherche d'une proximité élevée entre les données observées et les estimations fournies par le modèle



Sélection du modèle

- Objectif
  - Déterminer le modèle parcimonieux
    - Modèle réduit mais offrant un bon ajustement des données
- Constat
  - Nombre élevé de modèles possibles
    - Fonction du nombre de dimensions
    - Nombre de modèles log-linéaires pour k dimensions :

$$2^{2^k}$$

    - Exemple : Si k=3, alors 256 possibilités de modèles

Sélection du modèle

- **Calcul de la qualité de l'ajustement**
  - Déterminer le modèle qui fournit un bon ajustement des données
- Rapport de vraisemblance (*likelihood ratio*):

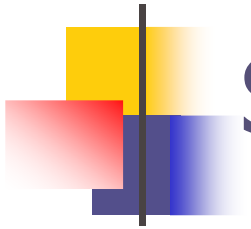
$$L^2 = \sum_{i1 i2 \ldots in} f_{i1 i2 \ldots in} * \log\left(\frac{f_{i1 i2 \ldots in}}{F_{i1 i2 \ldots in}}\right)$$

- L^2 suit une distribution du χ^2 avec un degré de liberté égal au nombre de paramètres λ égaux à zéro.
- Lorsque la mesure L^2 est grande en comparaison avec le degré de liberté, le modèle n'ajuste pas bien les données.

Sélection du modèle

$$L^2_{comparaison} = L^2_{\text{modèle}1} - L^2_{\text{modèle}2}$$

- Le modèle 1 doit être imbriqué dans le modèle 2 (posséder un ensemble d'interactions de moindre niveau).
- Le degré de liberté est calculé en soustrayant le degré de liberté du modèle 2 du degré de liberté du modèle 1.
- Lorsque $L^2_{comparaison}$ n'est pas significative, alors le modèle 1 est proche du modèle 2.
- **Conclusion**
  - Choisir le modèle qui a le moins de coefficients mais qui offre un bon ajustement



Sélection du modèle

- Approches
  - *Stepwise methods*
    - *Forward selection*
      - Ajouter les éléments un à un en privilégiant ceux qui aboutissent à une amélioration significative du test statistique.
    - *Backward elimination*
      - Débuter avec le modèle saturé et enlever les termes qui ont le moins d'impact sur la signification du test.
  - Variantes du *Backward elimination* (Atkins, Wemuth)

Découverte d'exceptions

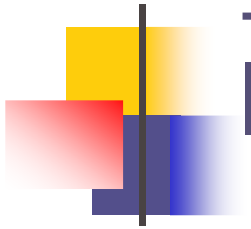
- Analyse des résidus
  - Permet la détection de valeurs extrêmes
    - par l'identification des cellules ayant un résidu élevé
  - Permet d'évaluer la qualité d'ajustement du modèle
    - Par l'identification des cellules où le modèle est le moins bien approprié (résidus les plus élevés) ou le mieux approprié (résidus les plus faibles)

$$résidu = \frac{fréquence_{observée} - fréquence_{espérée}}{\sqrt{fréquence_{espérée}}}$$

    - Si le modèle est approprié pour les données, les résidus des fréquences devraient consister en des valeurs positives et négatives d'une même amplitude et distribuées uniformément à travers les cellules du cube

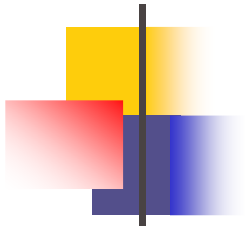
Avantages des modèles log-linéaires

- Modélisation de tableaux multidimensionnels et donc des cubes de données
- Prédiction des données
- Découverte d'exceptions



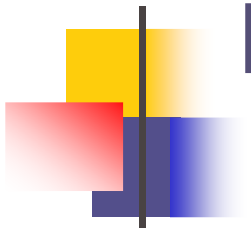
Inconvénients des modèles log-linéaires

- Applicabilité
  - Tableaux ayant une dimensionnalité faible ou moyenne (<15)
  - Tableaux (cubes) non creux
- Perte de précision
  - Due à l'approximation
  - Au niveau des valeurs extrêmes
- Échantillon de données
  - La taille doit dépasser 5 fois le nombre de cellules.



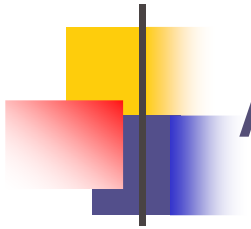
Nos travaux (en cours)

- Exploration des modèles log-linéaires pour
  - Le compactage des cubes de données
    - Identification du modèle le plus parcimonieux
    - Mise à jour incrémentale d'un modèle existant
    - Prise en compte de hiérarchies de dimensions
  - La détection des valeurs aberrantes
  - La prédiction de données
- Adaptation de nos travaux sur le DM
  - à la génération de règles d'association dans un contexte multidimensionnel
  - au regroupement de cellules, de dimensions ou de membres (modalités) de dimensions



Nos travaux

- Génération des motifs fermés
  - Objet (ou transaction) : fait du cube
  - Propriété (ou item) : modalité de dimension ou mesure
  - Algorithmes de construction du treillis de Galois ou utilisation d'un algorithme d'approximation du résultat
- Approximation du résultat (algo CIGA)
  - Limiter le nombre de motifs fermés
  - Éviter de conserver des motifs fermés fréquents inutiles pour les règles d'association
  - Gagner en temps d'exécution



Algorithme CIGA

- CIGA (*Closed Itemset Generation using an Automaton*)
- Part d'une matrice de cooccurrence entre items pour construire un automate
  - États : items
  - Transitions : cooccurences entre items dépassant un seuil de confiance donné
- Génère des motifs fermés par parcours de l'automate
- Inclut des opérations d'élagage

Matrice de cooccurrences d'items

	a	b	c	d	e	f	g	h
1	x		x	x	x	x		
2	x	x			x	x		
3	x		x				x	x
4					x		x	
5		x				x		
6	x	x	x					x
7		x		x	x		x	
8	x			x				
9		x	x	x		x		
10	x		x					

Contexte formel trié

	a	b	c	d	e	f	g	h
a	**6**							
b	2	**5**						
c	4	2	**5**					
d	2	2	2	**4**				
e	2	2	1	2	**4**			
f	2	***3***	2	2	2	**4**		
g	1	1	1	1	2	0	**3**	
h	2	1	2	0	0	0	1	**2**

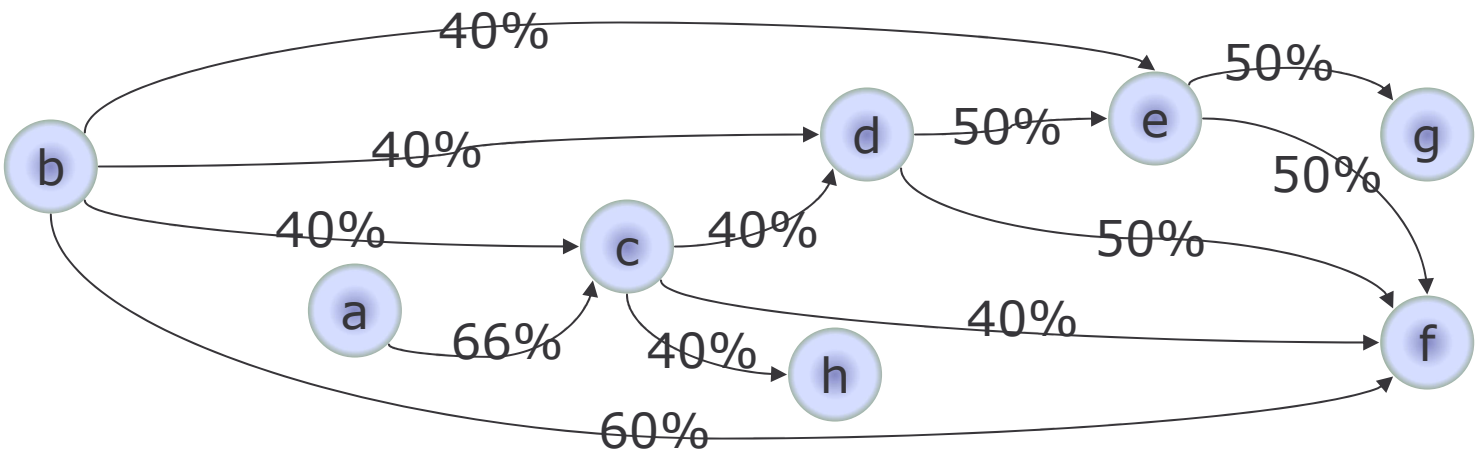
Matrice de cooccurrences

Exemple :
f apparaît 3 fois avec b
supp(b→f) = 3/10 = 30%
conf(b→f) = 3/5 = 60%

Paramètre de confiance

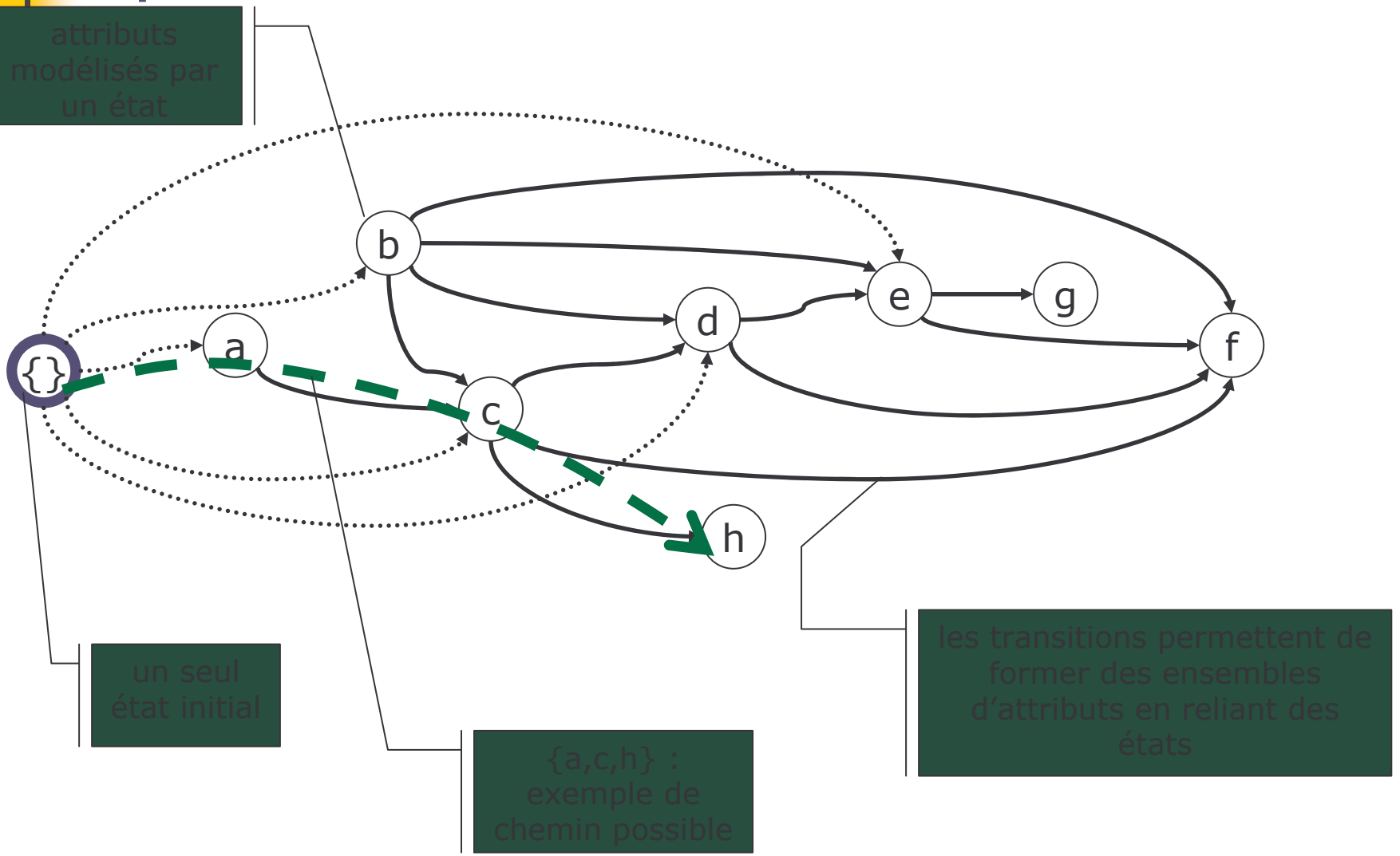
	{}	a	b	c	d	e	f	g	h
a	60								
b	50	33							
c	50	66	40						
d	40	33	40	40					
e	40	33	40		50				
f	40	33	60	40	50	50			
g						50			
h	20			40					

Avec une confiance minimale de 40%…



(Automate sans les points d'entrée)

Spécification de l'automate

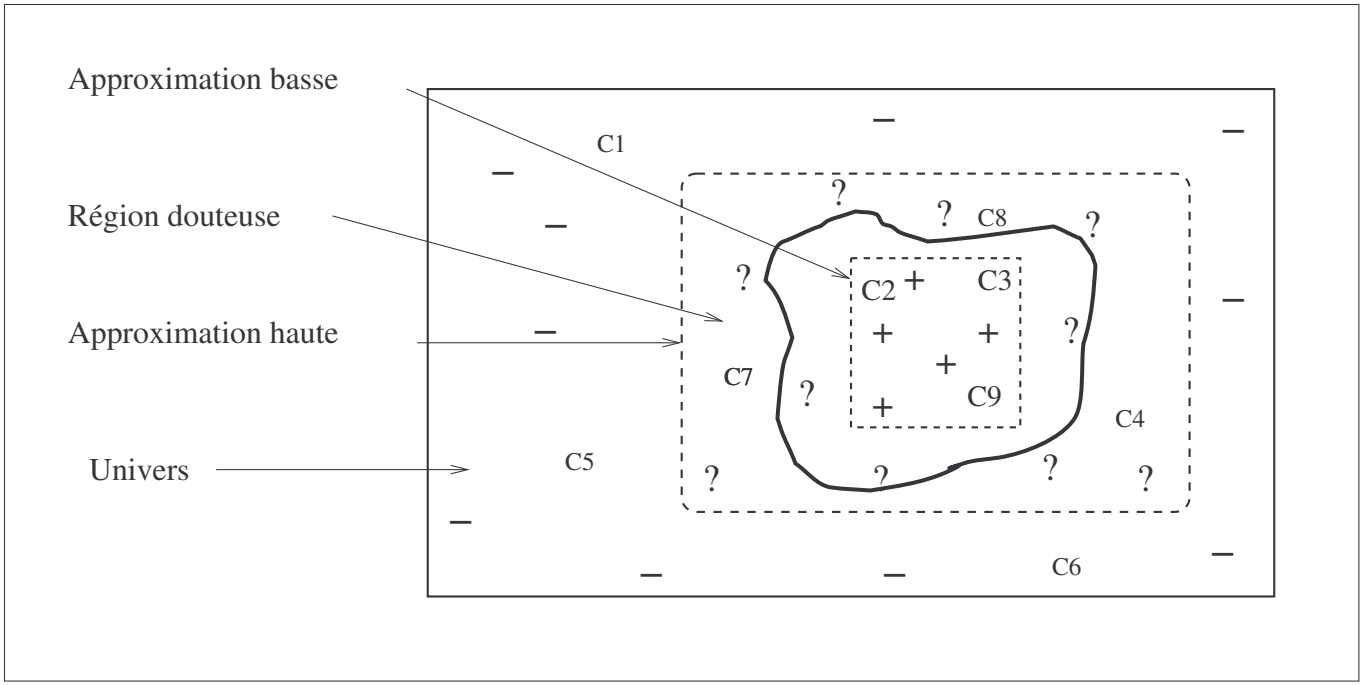


Nos travaux

- Appel à des techniques statistiques pour
  - la réduction de la dimensionnalité
    - Analyse en composantes principales, analyse en composantes indépendantes, ensembles d'approximation (*rough sets*)
- Utilisation de la théorie des « rough sets » pour
  - L'approximation des résultats d'une requête OLAP
  - La génération de règles de classification et de caractérisation.

Nos travaux

- Utilisation de la théorie des « rough sets »
  - Notions d'indiscernabilité et d'approximation



Nos travaux

- Utilisation de la théorie des « rough sets »
  - Règles de classification

$$(APDM = October \vee May) \wedge (ET = 90\ to\ 114\ months) \wedge (INVC = Belgium) \wedge (ATC = France) \longrightarrow (ICL = Textile)$$

$$(APDM = May \vee October) \wedge (ET = 65\ to\ 89\ months) \wedge (INVC = Canada) \wedge (ATC = France)) \vee ((APDM = May) \wedge (ET = 65\ to\ 89\ months) \wedge (INVC = Belgium) \wedge (ATC = France)) \longrightarrow (ICL = Pharmacy)$$

Nos travaux

- Appel à des techniques statistiques pour
  - la réduction de la dimensionnalité
    - Analyse en composantes principales, analyse en composantes indépendantes, ensembles d'approximation (*rough sets*)
  - La génération automatique de nouvelles hiérarchies de dimension
    - Classification hiérarchique
    - Algorithmes génétiques, ….
- Utilisation de la théorie des « rough sets » pour
  - L'approximation des résultats d'une requête OLAP
  - La génération de règles de classification et de caractérisation.

Autres travaux de recherche

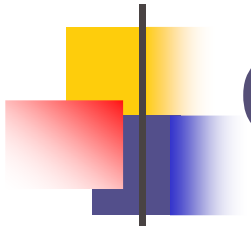
- Barbara et Wu
  - Utilisation des modèles log-linéaires pour le compactage des données et les requêtes approximatives
  - Traitement des cubes ayant des sections creuses et des exceptions
- Naouali
  - Nouveaux opérateurs OLAP; visualisation de cubes
  - Utilisation de la théorie des ensembles approximatifs pour l'identification des règles de classification et de caractérisation
  - Construction de grappes (*clusters*) de cellules
- A. Laurent
  - Fouille de données floues avec des opérateurs OLAP
  - Requêtes flexibles et traitement de données imparfaites; détection de cellules anormalement vides

Autres travaux de recherche

- Wu et al.
  - Étude des liens entre les dimensions et des associations entre des variables bimodales
  - Traitement de la forte dimensionnalité
- Lakshmanan et al.
  - Construction de cubes *quotient* (partitions de cubes)
  - Obtention de résumés sémantiques au sein des cubes
- Dong et al.
  - *Constrained gradient analysis*
  - Similaire à la notion de *cubegrade*
  - Prise en compte de contraintes
- Etc ...

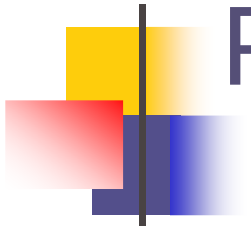
Autres travaux de recherche

- Wu et al.
  - Étude des liens entre les dimensions et des associations entre des variables bimodales
  - Traitement de la forte dimensionnalité
- Lakshmanan et al.
  - Construction de cubes *quotient* (partitions de cubes)
  - Obtention de résumés sémantiques au sein des cubes
- Dong et al.
  - *Constrained gradient analysis*
  - Similaire à la notion de *cubegrade*
  - Extraction de paires de cellules du cube qui sont différentes au niveau des mesures mais similaires au niveau des dimensions
  - Prise en compte de contraintes
- Etc ….



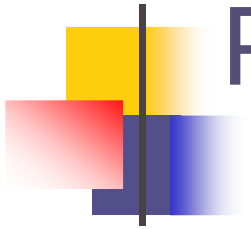
Conclusion

- La DCCD
  - Présente des défis intéressants dans un contexte multidimensionnel
  - Peut aboutir à des connaissances fort utiles décrivant des associations, des groupements ou des exceptions au niveau des données agrégées
  - Trouve des applications dans plusieurs domaines du monde réel



Références

- A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York, 1990.
- D. Barbara et X. Wu, Using Loglinear Models to Compress Datacubes, In *Proceedings of the International Workshop on Web-Age Information Management*, Shangai, Chine, juin 2000.
- R. Christensen, *Log-Linear Models and Logistic Regression*, Springer Verlag, 2nd edition, 1997.
- G. Dong, J. Han, J. M. W. Lam, J. Pei, K. Wang, W. Zou, Mining Constrained Gradients in Large Databases. *IEEE Trans. Knowl. Data Eng*. 16(8): 922-938 (2004).
- J. Han ¸et M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, Academic Press, États-Unis, 550 pages, 2001.
- T. Imielinski, L. Khachiyan, and A. Abdulghani. Cubegrades: Generalizing Association Rules. *Technical Report*, Rutgers University, August 2000.
- D. Knoke et P.J. Burke, *Log-Linear Models*, Sage Publications, Newberry Park, Cal., 1980.
- Laks V.S. Lakshmanan, Jian Pei, and Jiawei Han, Quotient Cube: How to Summarize the Semantics of a Data Cube, *Proc. Int. Conf. on Very Large Databases* (VLDB'02), Hong Kong, September 2002.
- A. Laurent. Bases de données multidimensionnelles floues et leur utilisation pour la fouille de données. Thèse de doctorat, Université Paris 6, septembre 2002.
- S. Naouali. Enrichissement d'entrepôts de données par la connaissance : application au Web. *Thèse de doctorat*, Université de Nantes, décembre 2004.



Références

- S. Nestorov et N. Jukic: Ad-Hoc Association-Rule Mining within the Data Warehouse. HICSS 2003.
- S. Sarawagi, R. Agrawal et N. Megiddo, *Discovery-Driven Exploration of OLAP Data Cubes*, *Proceedings of the Sixth Int'l Conference on Extending Database Technology* (EDBT), Valencia, Espagne, mars 1998.
- M. Welge. Knowledge Discovery from Databases. Talk, Nov. 2003.
- X. Wu, D. Barbara et Y. Yong, *Screening and Interpreting Multi-item Associations Based on Log-linear Modeling*, In Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., p. 276-285, août 2003.