Utilisation du calcul relationnel en sciences sociales : exemple de l'Observatoire des consommations alimentaires

Jean-Claude Poupa

INRA

LABORATOIRE DE RECHERCHE SUR LA CONSOMMATION

UTILISATION DU CALCUL RELATIONNEL EN SCIENCES SOCIALES : EXEMPLE DE L'OBSERVATOIRE DES CONSOMMATIONS ALIMENTAIRES

J.C. POUPA
INRA - Station d'économie
et sociologie rurales de Rennes

Novembre 1991

PRESENTATION POUR LE GROUPE DE RENOVATION
DE L'ENQUETE ALIMENTAIRE DE L'INSEE

UTILISATION DU CALCUL RELATIONNEL EN SCIENCES SOCIALES : EXEMPLE DE L'OBSERVATOIRE DES CONSOMMATIONS ALIMENTAIRES

1 - Représentation relationnelle des structures numériques

2 - Les opérations de calcul relationnel

3 - Le pilotage du processus d'exécution

4 - Les outils externes

5 - Conclusion : le projet DAVID

RESULTATS (1991)

Le logiciel INGRES permet d'administrer sur une station de travail d'équipe les grandes bases de données statistiques nationales et communautaires, traditionnellement gérées sur grands systèmes propriétaires.

Le moteur INGRES permet de concevoir un système homogène de gestion de bases de données adapté aux besoins scientifiques de traitement de l'information en sciences sociales. Un tel système reçoit les fichiers publiés par les institutions spécialisées, et produit des flots de vecteurs destinés aux logiciels de traitement statistique.

1 - REPRESENTATION RELATIONNELLE DES STRUCTURES NUMERIQUES

1.1. Représentation d'une matrice

- observation de *m* variables pour *n* ménages

  matrice **M** de type (*n*, *m*)

  -> tableau SAS

  -> R (i, j, m_{ij})

- table à 3 colonnes, appelées "attributs" de la relation. Les lignes sont appelées "tuples"ou "n_uplets"

i	j	m_{ij}
1	1	18
.	.	.
.	.	.

- le couple (i, j) identifie un terme m_{ij} et un seul

=> le couple (i, j) est la **clé primaire** de la relation R

- la relation R est un sous-ensemble du produit cartésien N X N X R

=> VERIFICATION INTEGRITE DE DOMAINE
ET INTEGRITE D'ENTITE

1.2. Représentation d'un arbre

- observation de la variable *k* pour l'individu *i* du ménage *m*

menage	individu	variable	valeur
m	i	k	x
.	.		.
.	.		.

- relation R (<u>menage</u>, <u>individu</u>, <u>variable</u>, valeur)
- sous-ensemble du produit cartésien NXNXNXR
- le triplet (menage, individu, variable) identifie une valeur et une seule. C'est la clé primaire de la relation R.
- le nombre d'occurrences d'une clé primaire est la **cardinalité** de la relation (= nombre de lignes de la table)

- observation des valeurs et des quantités d'un produit *p* inscrit à la ligne *l* d'un carnet de compte, pour le ménage *m* à la période *t*

menage	periode	produit	ligne	valeur	quantité
m	t	p	l	v	q
	.	.			.
	.	.			.

- relation R (<u>menage</u>, <u>periode</u>, <u>produit</u>, <u>ligne</u>, valeur, quantité)
- sous-ensemble du produit cartésien NxNxNxNxRxR
- le 4_uplet (menage, periode, produit, ligne) identifie une valeur et une seule. C'est la **clé primaire** de la relation R.

1.3. Représentation d'une nomenclature numérique

- nomenclature de produits évolutive dans le temps

produit	periode
p	t
.	.
.	.

- relation R (produit, periode)
- sous-ensemble du produit cartésien NxN
- le couple (produit, periode) est une clé primaire unique

- équivalences entre nomenclatures

produit	marche	e1	e2	e3	e4	e5	e6	INSEE	CIQUAL

- relation R (produit, marche, e1, e2, e3, e4, e5, e6, insee, ciqual)
- sous-ensemble de N^{10}
- l'attribut produit est une clé primaire unique, uniquement instrumentale
- le 7_uplet (marche, e1, e2, e3, e4, e5, e6) est le code produit secodip. C'est par construction une clé primaire unique.

1.4. Les vues

- Une vue est une relation virtuelle, générée dynamiquement à l'interrogation

- Exemples :
  - relation achat (menage, periode, produit, valeur, quantité)
  - relation dictionnaire (produit, marche, e1, e2, e3, e4, e5, e6, insee, ciqual)
  - vue secodip (menage, periode, marche, e1, e2, e3, e4, e5, e6, valeur, quantité)
  - vue insee (menage, periode, insee, valeur, quantite)
  - vue ciqual (menage, periode, ciqual, valeur, quantite)

- La relation dictionnaire est issue d'un ensemble d'opérations de calcul relationnel

- On peut créer des vues sur des vues

2- LES OPERATIONS DE CALCUL RELATIONNEL

LES OPERANDES SONT DES RELATIONS
L'OPERATION REND UNE RELATION

2.1. Restriction (unaire)

- sous-ensemble d'une relation pour lequel tous les n_uplets vérifient un prédicat de sélection (calcul des prédicats)

- **Exemple :** si le revenu est la variable numéro 10 dans la relation echantillon (menage, nv, x) la recherche des ménages ayant un revenu supérieur à 100 000 F est qualifiée par le prédicat (nv=10) et (x>100 000)

- rend une relation de cardinalité inférieure à la relation initiale

- requête fondamentale des processus d'interrogation

- les L4G construisent ces prédicats par interprétation du dialogue avec l'utilisateur

- SELECT menage, x from echantillon
  WHERE ((nv=10) and (x>100 000)) ;

2.2. Projection (unaire)

- relation obtenue en supprimant un ou plusieurs attributs (=> élimination des doublons)

- Opération fréquemment associée à l'utilisation de fonctions d'agrégation : dénombrement, somme, moyenne, maximum, minimum (GROUP BY)

- **Exemple 1 :** calcul du nombre d'individus par ménage

  - suppression des attributs variable et valeur r (menage, individu)

  - suppression de l'attribut individu, dénombrement pour chaque valeur du code menage, et introduction de l'attribut effectif
    => S (menage, effectif)

- **Exemple 2 :** calcul de la somme des achats par menage, periode et produit

  - suppression des attributs ligne, valeur et quantité et ajout de la somme des achats en valeur et en quantité pour chaque occurrence du triplet (menage, periode, produit)
    => (menage, periode, produit, valeur, quantité)

2.3. Union (binaire) INSERT

- union ensembliste de 2 sous-ensembles d'un même produit cartésien de domaines

- **Exemple 1 :** ajout du nombre d'individus par famille dans la relation (menage, nv, x)

• **Exemple 2 :** ajout du nombre d'achats dans la même relation

2.4. Différence (binaire) DELETE

- suppression dans le premier sous-ensemble des éléments présents dans le second sous-ensemble, les opérandes étant définis sur un même produit cartésien de domaines.

• **Exemple :** suppression d'une variable dans la relation R (menage, nv, x)

2.5. Produit cartésien (binaire)

- relation obtenue par concaténation de tous les attributs des relations opérandes

- le produit cartésien de 2 relations de cardinalités *n* et *m* donne une relation de cardinalité n x m

• **Exemple :** R(menage, produit, achat) x S(menage, poids)

m1, p1, a1, x1

m1, p1, a1, x2

200 000 x 5 000 => 10^9

- pour ajouter le poids, il faut exécuter une restriction : prédicat exprimant l'égalité des codes ménages dans les 2 relations

2.6. Jointure naturelle (binaire)

- Produit cartésien de 2 relations suivi d'une restriction sur le prédicat d'égalité des attributs de même nom, définis sur des domaines identiques

- R (menage, produit, achat) ⋈ S(menage, poids)

  predicat de jointure : R.menage = S.menage

  résultat T (menage, produit, achat, poids)

- Propriété : si la clé de jointure est clé primaire d'une relation opérande S, alors

  cardinalité (T) <= cardinalité (R)

2.7. Jointure naturelle d'ordre *n* (n_aire)

- opération associative et commutative

- application à la gestion de panels
  - pilote (menage, periode)
  - si *k* périodes

- *k* projections --> *k* relations

  P_k (menage)

- $P_1 \bowtie P_2 \bowtie .... \bowtie P_k$

  - coût maximal de l'ordre de

    $n\ (k \log_2 n + 1)$

  - temps d'exécution proportionnel à l'ordre de la jointure

- très utilisé en génération de vues

3- LE PILOTAGE DU PROCESSUS D'EXECUTION

3.1. Choix de l'organisation sur le disque

- Les composantes d'un vecteur doivent être physiquement contiguës sur le disque

- Les restructurations s'effectuent par ajout périodique d'un flot d'informations, sans qu'il y ait insertion d'informations élémentaires entre deux périodes : il n'y a pas lieu de réserver de l'espace pour les mises à jour (pages de débordement, réduction du taux de remplissage des pages...)

- La structure des relations d'ordre est généralement invariante pour un projet scientifique.

=>La structuration séquentielle indexée (ISAM) avec rangement séquentiel des données totalement indexées sur la clé primaire est suffisante. Elle garantit l'accès instantané à un vecteur (fonction de complexité en $\log_2$ n) et l'efficacité maximale pour un balayage de la table. L'espace disque consommé est le tiers de celui nécessaire avec une organisation arborescente classique.

3.2. Pilotage de la stratégie algorithmique

- **Exemples** :

  - secodip (menage, periode, marche, e1, e2, e3, e4, e5, e6, valeur, quantite)
    (marche = 250)

  - insee (menage, periode, insee, valeur, quantite)
    (insee/100 = 71)

- Vue qui exécute une jointure sur 2 relations
  achat (menage, periode, produit, valeur, quantité)
  et dictionnaire (produit, marche, e1, e2, e3, e4, e5, e6, insee, ciqual)

  - cardinalité de achat 2×10^6
  - cardinalité de dictionnaire 20 000

- Demande : extraction de vecteurs sur l'espace de l'échantillon des ménages

- Stratégie souhaitée et exécutée

  - fichier associé à la relation achat rangé dans l'ordre croissant des triplets

    (produit, periode, menage)

    => recherche sur les index ISAM et rapatriement d'un flot

    => inutile de créer un index sur achat, la structuration ISAM fournissant un index sans coût supplémentaire

  - fichier associé à la relation dictionnaire résidant en mémoire principale

- Stratégie inadaptée

  - indexation ISAM dans l'ordre croissant
    (menage, periode, produit)

  - nécessité de consulter tous les ménages

  - INGRES refuse et exécute un réordonnancement de la table dans l'espace temporaire, (=> coût en $n \log_2 n$) pour revenir au cas précédent.

- Stratégie ORACLE V5

  - le système connait la relation d'ordre seulement si un index a été explicitement créé par un ordre SQL CREATE INDEX

  - Un index est une relation dont la taille est de même ordre de grandeur que la relation indexée

  - Si l'index ORACLE existe, il sera ignoré pour l'interrogation par la vue insee, l'optimiseur syntaxique en décidant ainsi :

=> BALAYAGE DE L'ARBORESCENCE POUR CHAQUE FROMAGE
(--> COUT EN k x n/2, SI k FROMAGES)
k = 884 => 1H15 MN

3.3. Optimisation statistique : principe

--> **Exemple 1 :** introduction du poids du menage dans la relation achat

- Achat (menage, produit, periode,
valeur, quantité) ⋈ poids (menage, poids)

- Stratégie 1

  - pour chaque n_uplet de achat, restriction sur poids pour extraire le poids du ménage

=> - lecture séquentielle du fichier image de achat
- le fichier image de poids reste en mémoire

- Stratégie 2

  - pour chaque n_uplet de poids, restriction sur achat pour extraire les ménages concernés

  => multiplication des accès disques pour retrouver les ménages éparpillés sur le disque

- Pour choisir l'algorithme efficace, l'optimisation syntaxique ne permet pas de décider

--> **Exemple 2 :** extraction de la consommation d'un nutriment

répertoire (produit, nutriment, valeur)

achat (menage, periode, produit, quantite)

10 000 produits }
}=> 10^6 n_uplets
1 000 nutriments }

2×10^6 achats

- La requête exécute une jointure naturelle répertoire ⋈ achat avec une double restriction (nutriment = *n*) et (periode > 66)

- L'optimisation syntaxique ne permet pas de décider le choix de l'algorithme

- Optimisation statistique

  restriction (nutriment = *n*) => relation de cardinalité 10 000

  restriction (periode > 66) => relation de cardinalité 1.5×10^6

- Si base répartie, la base d'accueil de la relation achat importe le résultat de la restriction de la base repertoire

3.4. Contrôle des ressources

- L'administrateur peut visualiser l'algorithme d'exécution choisi et les coûts associés aux opérations (majorants du nombre d'entrées-sorties et du nombre d'instructions à exécuter)

- Un mécanisme de règles permet de contrôler les ressources et de ne pas autoriser l'exécution de requêtes trop coûteuses. Ces informations sont gérées dans des relations.

3.5. Optimisation de l'utilisation des ressources

- Nomenclature secodip : 20 000 produits
  remplacement du 7_uplet (marche, e1, e2, e3, e4, e5, e6) par un attribut instrumental

- Méthode

  - création de la relation secodip (marche, e1, e2, e3, e4, e5, e6)
  - modification du modèle physique en séquentiel ordonné
  - création du fichier image de la relation
  - exécution d'un programme externe (sequence) qui affecte un numéro d'ordre à chaque ligne du fichier image de la relation
  - transformation du fichier créé en relation dictionnaire (produit, marche, e1, e2, e3, e4, e5, e6)
  - modification du modèle physique en ISAM sur la clé primaire

- Choix des domaines

  - ajustement précis des types d'entiers
    (sur 1,2 ou 4 octets)

  - remplacement des numéros de ménages par un numéro instrumental
    (=> secret statistique)

  - remplacement d'un groupe d'attributs par un code instrumental

  - création de vues

4- LES OUTILS EXTERNES

- Création d'un ensemble de procédures sous UNIX externes à INGRES
  - séquence
  - dtag (distinction des achats identiques)
  - vectoriser
  - interface LEDA
  - interface SECODIP
  - interface EUROSTAT

- Intégration prévue de procédures dans le moteur INGRES, sous forme de nouveaux domaines de définition, après validation.

- Création de procédures SQL pour l'enchaînement des opérations

- Le système INGRES intègre les outils créés pour les besoins scientifiques. L'utilisation du précompilateur (accès à la base depuis un programme L3G) ne se justifie plus.

4.1. La fonction de vectorisation

- C'est une fonction F (A, B, C) qui rend une matrice M (tableau SAS), A, B, et C étant 3 relations construites sous INGRES

  - A décrit les lignes A (menage, periode)

  - B décrit les colonnes B (produit)

  - C décrit une relation ayant pour clé primaire le produit cartésien A X B
    C (menage, periode, produit, valeur, quantite)

- Les relations A, B et C sont construites au terme d'une suite d'opérations relationnelles, avec contrôle des définitions et vérification des cardinalités

- Le modèle physique de ces relations est transformé en séquentiel ordonné, avec les relations d'ordre adaptées

- Les fichiers images des relations sont édités

- Le programme "vectoriser" lit ces fichiers et édite un fichier image classique d'un tableau SAS

4.2. L'interface leda

- La procédure interprète le flot binaire associé à l'arborescence leda et rend un ensemble de fichiers images de relations

- La lecture des fichiers en provenance des systèmes propriétaires, transcrits sur bandes magnétiques, s'effectue avec un programme diffusé par l'AFFU : ansiread

- Les primitives système de décodage des types de données élémentaires (reconnaissance du format variable, conversion ebcdic/ascii ; conversion des représentations binaires des nombres entiers en flottants) sont développées à l'INRA

- La procédure effectue une première reconnaissance syntaxique de la suite d'arborescences et édite le fichier image d'un dictionnaire des niveaux présents, en précisant la longueur et le nombre d'occurences.

- Ce fichier est repris sous ingres et la relation résultat est utilisée pour construire un dictionnaire des variables

- Des attributs spécifiques du dictionnaire des variables permettent de choisir ces variables, leur type (chaine de caractères de longueur x ou valeurs numériques) et de paramétrer l'édition en fonction de relations cibles, avec clés primaires instrumentales

- La procédure reconnait tous les fichiers leda, créés sous IBM ou BULL

- La procédure permet de créer rapidement une base INGRES contenant l'intégralité de l'information sans connaissance du dictionnaire leda

5- CONCLUSION : LE PROJET DAVID

- Hypothèses sur l'informatique des prochaines années

  - Reseau et bases de données
  - Down sizing
  - Approche objet
  - Langages graphiques
  - SGBD répartis hétérogènes

Stratégie DAVID

  - élaboration des composants (SQL, langage C)
  - validation des protocoles
  - portage vers les SGBD du marché et intégration du code L3G dans les moteurs SQL
  - création des interfaces utilisateurs en environnement graphique "client"
  - interconnexion des bases d'équipes sur un réseau national et système de bases réparties

- Poursuite des développements avec INGRES dans l'état actuel de l'offre du marché

Distribution

Archivage

Vectorisation

Importation de Données