Contribution de la classification automatique pour l'organisation et l'interrogation d'un corpus de "Petites annonces"

Philippe Peter

INRIA

CENTRE DE RENNES

IRISA

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

Rapports de Recherche

N° 276

CONTRIBUTION DE LA CLASSIFICATION AUTOMATIQUE POUR L'ORGANISATION ET L'INTERROGATION D'UN CORPUS DE "PETITES ANNONCES"

Philippe PETER

Mars 1984

IRISA

INSTITUT DE RECHERCHE EN INFORMATIQUE
ET SYSTÈMES ALÉATOIRES

Campus Universitaire de Beaulieu
Avenue du Général Leclerc
35042 - RENNES CÉDEX
FRANCE
Tél. : (99) 36.20.00
Télex : UNIRISA 95 0473 F

CONTRIBUTION DE LA CLASSIFICATION AUTOMATIQUE POUR L'ORGANISATION ET L'INTERROGATION D'UN CORPUS DE "PETITES ANNONCES"

Ph. PETER

Publication Interne n° 218
32 pages

Résumé : On se propose de montrer comment la classification des données peut aider à l'élaboration d'un système de traitement automatique de petites annonces immobilières. Le but d'un tel système étant l'affectation d'une classe de taille "raisonnable" de petites annonces "complémentaires" à une annonce proposée par un utilisateur.
Une phase préliminaire utilisant la même approche permet la "compréhension" du corpus des données.
On rappelle sommairement les techniques de classification employées.
On finira en faisant le point sur les différentes perspectives d'amélioration du système à court et moyen termes.

Abstract : In this paper, we show how clustering analysis can help to work out an automatic processing of classified advertissements.
A such system delivers the classes of "complementary" ads to an user who asked about his ad.
A preliminary phase using the same approach allows us to "understand" the data corpus. The technics of automatic classification used are shortly realled. Finally we'll at the different possible improvements which may be considered at short term.

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
(L.A. 227)
UNIVERSITÉ DE RENNES 1 I.N.S.A. DE RENNES

INSTITUT NATIONAL DE RECHERCHE
EN INFORMATIQUE ET EN AUTOMATIQUE
(LABORATOIRE DE RENNES)

1 - METHODES DE CLASSIFICATION EMPLOYEES :

1-1 Généralités :

Le point de départ de la classification est le croisement de l'ensemble des annonces par un ensemble de variables descriptives. Deux classifications sont alors possibles :
- celle des annonces qui permet de partitionner leur ensemble en classes ;
- celle des variables descriptives qui rend compte des principales tendances du comportement de l'échantillon.

1-2 Variable descriptive choisie :

Nous voulions représenter les annonces simplement, mais aussi coder des choses aussi différentes que la présence ou l'absence d'une caractérisque, que des valeurs numériques (codées éventuellement par intervalles) tels que des prix ou des surfaces par exemple. L'attribut de description nous a alors semblé adapter à nos besoins. Nous le définirons au moyen d'une application a de E (l'ensemble des annonces) dans {0,1}.

a : E |---> {0,1}
x ---> a(x) = 1 si x possède l'attribut a,
= 0 sinon

Nous appelerons Ea le sous-ensemble de E composé de toutes les annonces possédant l'attribut a (Ea=a-1(E)) et poseront : n(a) = card(Ea) et n = card(E).

1-3 Méthodes de classification utilisées pour les attributs :

1-4-1 Algorithme de la Vraisemblance de Lien [3], [4] :

1-3-1-1 Indice de proximité :

Soient a et b deux attributs et Ea et Eb les sous-ensembles d'annonces qu'ils définissent. L'indice de proximité p(a,b) sera basé sur la statistique s=card(Ea $\cap$ Eb), c'est à dire le nombre d'annonces possédant à la fois l'attribut a et l'attribut b. Il s'avère cependant nécessaire d'introduire une Hypothèse d'Absence de Lien N pour neutraliser un "effet de taille" (si a et b sont fréquents (resp. rares), s sera grand (resp. petit)).

PLAN
====

0 - INTRODUCTION :

Une équipe de l'I.R.I.S.A. s'intéressant à la conception d'un système de traitement automatique de petites annonces immobilières [7], [8], il nous a semblé intéressant d'en étudier statistiquement un corpus afin de mieux connaître les principales tendances du marché et éventuellement d'aider le système à sélectionner des annonces complémentaires à une annonce donnée (pouvant répondre au besoin de l'annonceur).

Dans cet article, nous ne présenterons que très brièvement les méthodes de classification employées laissant le soin aux lecteurs intéressés de consulter les publications citées en référence. Nous analyserons attentivement la classification des attributs et partitionnerons leur ensemble en vue de la compréhension du marché actuel des petites annonces. Nous déduirons de la classification des annonces des classes potentielles d'annonces complémentaires, ce qui permettra l'écriture d'un petit système.

1 - METHODES DE CLASSIFICATION EMPLOYEES :

1-1 Généralités :

Le point de départ de la classification est le croisement de l'ensemble des annonces par un ensemble de variables descriptives. Deux classifications sont alors possibles :
- celle des annonces qui permet de partitionner leur ensemble en classes ;
- celle des variables descriptives qui rend compte des principales tendances du comportement de l'échantillon.

1-2 Variable descriptive choisie :

Nous voulions représenter les annonces simplement, mais aussi coder des choses aussi différentes que la présence ou l'absence d'une caractérisque, que des valeurs numériques (codées éventuellement par intervalles) tels que des prix ou des surfaces par exemple. L'attribut de description nous a alors semblé adapter à nos besoins. Nous le définirons au moyen d'une application a de E (l'ensemble des annonces) dans {0,1}.

a : E |---> {0,1}
x ---> a(x) = 1 si x possède l'attribut a,
= 0 sinon

Nous appelerons Ea le sous-ensemble de E composé de toutes les annonces possédant l'attribut a (Ea=a-1(E)) et poseront : n(a) = card(Ea) et n = card(E).

1-3 Méthodes de classification utilisées pour les attributs :

1-4-1 Algorithme de la Vraisemblance de Lien [3], [4] :

1-3-1-1 Indice de proximité :

Soient a et b deux attributs et Ea et Eb les sous-ensembles d'annonces qu'ils définissent. L'indice de proximité p(a,b) sera basé sur la statistique s=card(Ea $\cap$ Eb), c'est à dire le nombre d'annonces possédant à la fois l'attribut a et l'attribut b. Il s'avère cependant nécessaire d'introduire une Hypothèse d'Absence de Lien N pour neutraliser un "effet de taille" (si a et b sont fréquents (resp. rares), s sera grand (resp. petit)).

Pour ce faire associons à (Ea,Eb), un couple d'éléments aléatoires (X,Y) tel que X corresponde à Ea et Y corresponde à Eb avec card(X) = card(Ea) et card(Y) = card(Eb), et considérons la variable aléatoire S = card(X,Y). L'indice de proximité choisi est alors p(a,b)=PrN(S<s).

Dans le cas de l'hypothèse d'absence de lien N3 (correspondant à un modèle poissonnien et favorisant les associations entre attributs rares), nous aurons l'indice suivant :

$$p(a,b) = \frac{s-(n(a)*n(b)/n)}{(n(a)*n(b)/n)^{1/2}}$$

1-3-1-2 Démarche générale de l'algorithme :

(a) calcul de l'indice de proximité pour tout couple d'attribut ;

(b) tout attribut forme une classe ;

(c) recherche des couples de classes (n) les plus proches (Ai et Bi)

(d) pour i=1 à n faire ;
  (d1) formation de la classe Ci=AiUBi
  (d2) destruction des classes Ai et Bi
  (d3) réactualisation de la table des proximités

(e) si nombre de classe > 1 alors aller à (c)

(f) stop .

Cet algorithme implémenté au C.I.C.B. [5], conduit à la représentation polonaise de l'arbre de classification.

1-3-2 Algorithme basé sur la corrélation (A.B.C.) [6] :

1-3-2-1 Indice de proximité :

L'idée consiste à considérer une colonne j (un attribut) comme une variable numérique Xj. L'indice de proximité entre les colonnes j et h sera alors le coefficient de corrélation entre les variables Xj et Xh.

1-3-2-2 Démarche générale de l'algorithme :

Elle est analogue à celle de l'Algorithme de la Vraisemblannce de Lien (A.V.L.).

1-4 Méthode de classification pour les annonces :

Pour une raison de temps calcul, seul l'algorithme de la vraisemblance de lien a été utilisé. La méthode est totalement symétrique à celle employée pour la classification des attributs.

1-5 Aides à l'interprétation des résultats :

1-5-1 statistique globale des niveaux :

Elle mesure pour chaque niveau de l'arbre la cohésion des classes formées.

1-5-2 statistique locale des niveaux :

Sa valeur augmente lorsqu'une classe en cours de formation se confirme et tombe quand cette classe est formée (cohérente). Les maximums locaux de cette statistique correspondent donc à des niveaux d'achèvement de classes. Les noeuds de l'arbre correspondants seront considérés comme significatifs. L'arbre de classification sera donc condensé à ces noeuds significatifs.

1-5-3 Indice de dispersion d'un élément :

Plus cet indice est grand, plus l'élément (ici une annonce) entraîne sa classe.

2 - CONSTITUTION ET CODAGE DE L'ECHANTILLON :

2-1 Constitution :

Mille petites annonces immobilières recopiées dans les journaux "Ouest-France", "Rennes Pub" et "Le 35" ont été mises sur fichier. Nous avons délibérément exclu les annonces portant sur les fonds de commerce et sur les résidences de vacances. La plupart de ces annonces émanent d'agences immobilières ou de notaires: la description de l'objet de la transaction est souvent très brève. Il est également à noter que le marché actuel des petites annonces est très particulier; en effet près de 88% d'entre-elles sont des offres de vente, viennent ensuite (dans notre échantillon) les offres de location (8,2%), les demandes de location (2,4%), les demandes d'achat (1,3%) et les viagers (quasi-inexistants -un pour mille). Ce marché doit cependant être sujet à des fluctuations saisonnières, les demandes de location devant être beaucoup plus fréquentes au moment de la rentrée universitaire par exemple.

2-2 Choix des attributs :

Une étude faite au préalable [2] a servi de point de départ. Deux cent trente huit attributs avaient alors été retenus, ceci en collaboration avec l'agence Havas. Ils se sont avérés beaucoup trop nombreux pour former des classes bien typées. Notre objectif était de ramener leur nombre autour de la centaine. Pour ce faire, nous avons opéré principalement de trois manières:

* élimination des synonymes :
  Par exemple un seul attribut pour T3, F3 et 3 pièces

* élimination des attributs trop rares :
  Le seuil a été fixé à dix apparitions pour mille annonces

* redéfinition des attributs de lieu:
  Leur nombre avoisinait quatre vingts soit un tiers du total et on pouvait y reconnaître bon nombre de communes du département et de quartiers de Rennes. Nous nous sommes limités à un découpage en zones pour Rennes (nord, sud, ...)et en distance par rapport à Rennes pour les environs (moins de quinze kilomètres, ...).

Ceci nous a conduit à ne retenir que quatre vingt dix sept attributs (liste en annexe).

2-3 Construction du fichier de données :

Contrairement à ce qui avait été fait lors de l'étude précédente, le fichier a été automatiquement généré par programme. La technique employée a été d'isoler les mots (chaines de caractères) de leur contexte et de les comparer un par un aux attributs retenus. Il a cependant été parfois nécessaire de tenir compte du contexte, essentiellement dans les deux cas suivants:

- les nombres;
- les expressions composées
  (par exemple cherche ... à louer)

3 - CLASSIFICATION DES ATTRIBUTS :

A) Classification basée sur A.B.C. :

L'arbre obtenu comporte vingt neuf noeuds significatifs pour quatre vingt six niveaux. Le maximum de la statistique globale est atteint au niveau soixante treize avec 16,912. Pour obtenir des classes plus cohérentes et moins d'attributs neutres, l'arbre a été coupé au niveau soixante quinze qui est d'ailleurs un maximum de la statistique locale. Il se dégage huit classes significatives. Pour chacune d'entre-elles, la liste des attributs est donnée par valeurs de dispersions décroissantes.

Première classe :

- formée au niveau 52

- 7 attributs :
  à vendre, Rennes centre, studio, une pièce, moins de 200000frs, moins de 50m2, restauré

- cette classe très significative, définit parfaitement les très petits logements à vendre du centre ville. Il s'agit de F1 ou de studios éventellement restaurés.

Deuxième classe :

- formée au niveau 68

- 12 attributs :
  appartement, quatre pièces, divers Rennes, trois pièces, Rennes sud, deux pièces, premier ou deuxième étage, 71-90m2, Rennes nord, 51-70m2, prêt, Rennes ouest.

- on retrouve ici les appartements moyens de deux à quatre pièces. Cette classe est représentative de la formulation des offres de vente d'appartements par les agences immobilières où seuls sont indiqués : le type du logement, le lieu, la surface et une éventuelle possibilité de prêt. Ces appartements recouvrent toute la ville.

Troisième classe :

- formée au niveau 71

- 16 attributs :
  à louer, chambre, meublé, région de Saint-Malo, moins de 1000frs, cherche à louer, local, plus de sept pièces, plus de 2000frs, 1001 à 2000frs, entre 15 et 40km de Rennes, achète, particulier, plus de 600000frs, neuf, indépendant.

- cette classe est plus hétérogène. Le marché de l'immobilier actuel comporte 88% d'offres de vente : cette classe regroupe le reste (achats, offres et demandes de location)
On peut distinguer certains groupements associant des attributs à dispersion relativement élevée :
* cherche à louer, chambre, meublé (noeud 7)
* à louer, région de Saint-Malo, <loyer> (noeud 11)
On s'aperçoit que le montant des loyers n'est pas significatif.

Principaux types regroupés dans cette classe :

- demandes de location de chambres ou de meublés
- offres de location à la mer (région de Saint-Malo)
- offres de location de locaux
- demandes d'achat de grosses propriétés

Quatrième classe :

- formée au niveau 67

- 7 attributs :
  cuisine, cheminée, séjour, aménagé, grand, salle de bain, terrasse.

- on ne retrouve plus ici de caractéristiques formelles de logements. Cette classe regroupe des éléments de confort intérieur avec l'énumération des pièces. Ces attributs s'appliquent plus aux maisons qu'aux appartements.

Cinquième classe :

- formée au niveau 73

- 21 attributs :
  maison, terrain, jardin, autre lieu, sous-sol, six ou sept pièces, 501 à 1000m2, moins de 15km de Rennes, 151 à 500m2, construtible, garage, 500001 à 600000frs, campagne, 1001 à 2000m2, 2001 à 3000m2, à restaurer, plus de 3000m2, dépendance, cellier, 111 à 150m2, Rennes sud-est.

- malgré un nombre important d'attributs, cette classe est très cohérente. On observe deux noyaux principaux d'attributs à forte dispersion :
* r<=15km, maison, jardin, 6-7 pièces, sous-sol

* autre lieu, terrain, constructible

En regardant de plus près l'arbre de classification, on se rend compte que cette classe se divise en deux au niveau 65. La première sous-classe décrit les maisons "classiques" de la banlieue de Rennes (moins de quinze kilomètres), de 500001 à 600000 francs avec six ou sept pièces et bénéficiant d'un jardin et d'un garage. La seconde sous-classe associe les terrains constructibles situés beaucoup plus loin.

La réunion de ces deux sous-classes peut s'expliquer par le fait qu'une grande partie des maisons "classiques" de banlieue soient entourées de terrain.

Sixième classe :

- formée au niveau 64

- 6 attributs :
  cinq pièces, pavillon, 91 à 110m2, 300001 à 500000frs, plain-pied, frais.

- cette classe est également très typée. On y reconnaît les pavillons préfabriqués, de plain-pied, peu chers, vite construits et vendus "clefs en mains" par le promoteur.

Septième classe :

- formée au niveau 74

- 17 attributs :
  chauffage, individuel, cave, gaz, séchoir, parking, 20001 à 300000frs, ascenseur, en pierres, troisième ou quatrième étage, confort, grenier, libre, balcon, au delà du quatrième étage, Rennes sud-ouest, clos.

- tout comme la quatrième classe, celle-ci ne définit pas vraiment un type d'habitation mais décrit les facilités et l'environnement du logement.
Ces attributs sont plus orientés vers les appartements que vers les maisons.

Noyaux d'attributs à dispersion relativement élevée :
* parking, cave, séchoir, 200001 à 300000frs
* chauffage, individuel, gaz

Cette classe englobe des appartements décrits moins succintement que dans une formulation du type de celles des agences immobilières.

Huitième classe :

- formée au niveau 35

- 4 attributs :
  petit, collectif, immeuble, récent.

- ces quatre attributs définissent l'environnement de l'appartement.

Attributs neutres :

Ils sont au nombre de sept :
Rennes est, rez de chaussée, standing, bon état, calme, beau, intéressant.

Ils ne rejoignent des classes qu'à un niveau très élevé de l'arbre et sont de dispersion très faible.

Remarques générales :

* les surfaces de terrain, les loyers et les quartiers de Rennes en général ne sont pas très significatifs.

* association quasi-immédiate entre le nombre de pièces et la surface habitable (environ 20m2 par pièce).

Récapitulatif :

Première classe : logements d'une pièce au centre ville ;

Deuxième classe : appartements moyens décrits par les agences immobilières ;

Troisième classe : transactions autres qu'une offre de vente ;

Quatrième classe : éléments de confort intérieur (orientés maison) ;

Cinquième classe : - maisons classiques,
- terrains ;

Sixième classe : pavillons préfabriqués ;

Septième classe : facilités et environnement du logement (orientés appartement) ;

Huitième classe : petit collectif, immeuble récent.

B) Classification basée sur A.V.L. :

L'arbre obtenu comporte vingt noeuds significatifs pour quatre vingt sept niveaux. On atteint la plus forte statistique globale au niveau cinquante trois avec 16,617. En coupant l'arbre à ce niveau, beaucoup d'attributs se retrouveraient isolés, aussi va-t-on le couper au niveau soixante quatorze (statistique globale : 15,513). On distingue ici huit classes. Comme précédemment, pour chacune d'entre-elles, les attributs seront énumérés par valeurs de dispersion décroissantes. On regardera également ce qu'il advient de ces classes en coupant leur sous-arbre respectif au niveau cinquante trois.

Première classe :

- formée au niveau 52

- 5 attributs :
  Rennes centre, une pièce, moins de 200000frs, studio, moins de 50m2.

- elle décrit les studios et F1 du centre ville.

Deuxième classe :

- formée au niveau 73

- 5 attributs :
  plus de sept pièces, plus de 600000frs, terrasse, neuf, beau.

- elle caractérise les grande propriétés, mais les attributs qui la compose sont peu liés entre eux : au niveau cinquante trois, la classe est totalement désagrégée.

Troisième classe :

- formée au niveau 75

- 32 attributs :
  maison, terrain, jardin, à louer, cheminée, séjour, sous-sol, cuisine, autre lieu, 501 à 1000m2, six ou sept pièces, aménagé, moins de 15km de Rennes, 151 à 500m2, 500001 à 600000frs, garage, salle de bain, région de Saint-Malo, grand, constructible, moins de 1000frs, local, 2001 à 3000m2, à restaurer, plus de 2000frs, dépendance, plus de 3000m2, de 1001 à 2000frs, Rennes sud-est, rez de chaussée, de 111 à 150m2.

- on peut y reconnaître des descriptions de terrains et de maisons relativement importantes un peu en dehors de Rennes ainsi que des offres de location. Pour y voir plus clair, on va couper le sous-arbre au niveau cinquante trois. Quatre sous-classes sont alors obtenues.

a) cheminée, séjour, cuisine, aménagé, salle de bain, grand, 500001 à 600000frs, Rennes sud-est.
Exceptés les deux derniers attributs, on trouve ici un noyau à forte dispersion (niveau 33).
cette sous-classe caractérise des maisons assez luxueuses.

b) à louer, région de Saint-Malo, moins de 1000frs, local, plus de 2000frs, 1001 à 2000frs.
Cette sous-classe regroupe les offres de location qui sont de deux types :
- locations de locaux,
- locations dans la région de Saint-Malo.

c) terrain, autre lieu, 501 à 1000m2, 151 à 500m2, constructible, 1001 à 2000m2, 2001 à 3000m2, plus de 3000m2.
On trouve là les terrains qui sont souvent loin de Rennes.

d) maison, jardin, sous-sol, six ou sept pièces, moins de 15km de Rennes, garage, 111 à 150m2.
Cette sous-classe décrit les maisons classiques avec jardin, garage, six ou sept pièces dans les environs de Rennes.

Quatrième classe :

- formée au niveau 57

- 5 attributs :
chambre, meublé, cherche à louer, particulier, achète.

- cette classe comprend les demandes de location de chambres ou meublés ainsi que les demandes d'achat.
Au niveau cinquante trois, il ne reste que le noyau : chambre, meublé, cherche à louer.

Cinquième classe :

- formée au niveau 70

- 5 attributs :
quatre pièces, Rennes sud, 71 à 90m2, Rennes nord, prêt.

- quatre pièces est l'élément entraînant de cette classe peu liée qui éclate lorsqu'on la coupe au niveau cinquante trois.

Sixième classe :

- formée au niveau 63

- 18 attributs :
cave, collectif, petit, séchoir, 200001 à 300000frs, parking, balcon, campagne, en pierres, troisième ou quatrième étage, confort, grenier, immeuble, Rennes sud-ouest, cellier entre 15 et 40km de Rennes, clos, récent.

- cette classe est difficilement interprétable telle quelle. On va de suite couper le sous-arbre correspondant au niveau cinquante trois ; on se trouve alors en présence de trois sous-classes :

a) cave, séchoir, parking, balcon, 200001 à 300000frs, troisième ou quatrième étage, Rennes sud-ouest.
On remarque au niveau huit la réunion de trois attributs à forte dispersion : cave, séchoir, parking.
Cette sous-classe caractérise de petits appartements confortables.

b) campagne, grenier, entre 15 et 40km de Rennes, clos.
Cette sous-classe définit les maisons de campagne.

c) collectif, petit, immeuble, récent.
On situe ici le cadre du logement.

Septième classe :

- formée au niveau 69

- 10 attributs :
chauffage, individuel, appartement, gaz, premier ou deuxième étage, trois pièces, ascenseur, 51 à 70m2, au dessus du quatrième étage, libre.

- cette classe définit les appartements moyens. Au niveau cinquante trois, elle se divise en deux :
d'un côté la description formelle du logement, de l'autre les éléments de confort.

Huitième classe :

- formée au niveau 74

- 9 attributs :
  cinq pièces, pavillon, 300001 à 500000frs, 91 à 110m2, plain-pied, frais, calme, standing, bon état.

- cette classe caractérise les pavillons préfabriqués.
Au niveau cinquante trois, on observe deux sous-classes principales :
dans l'une la désignation et le prix, dans l'autre le nombre de pièces et la surface.

Attributs neutres :

Huit attributs :
divers Rennes, deux pièces, à vendre, Rennes ouest, Rennes est, indépendant, restauré, intéressant.

Remarques :
---

* les loyers et les surfaces de terrains ne sont pas significatifs ; c'est la même chose pour les quartiers de Rennes en général.

* on observe une rapide association entre le nombre de pièces et la surface habitable.

Récapitulatif :

Première classe : logements d'une pièce au centre ville

Deuxième classe : très grande propriétés ;

Troisième classe : - maisons luxueuses,
- maisons classiques,
- terrains,
- offres de location ;

Quatrième classe : - demandes de location de chambre,
- demandes d'achat ;

Cinquième classe : Logements de quatre pièces ;

Sixième classe : - petits appartements confortables,
- propriétés à la campagne,
- petits collectifs ;

Septième classe : appartements moyens ;

Huitième classe : pavillons préfabriqués.

C) Comparaison des deux classifications :

1) Coupure de l'arbre :

La coupure de l'arbre intervient sensiblement au même niveau. Il s'avère cependant nécessaire avec A.V.L. de subdiviser certaines classes pour une meilleure compréhension.

2) Comportement des classes au niveaux inférieurs :

* Corrélation : généralement les classes se subdivisent en sous-classes (n'éclatent pas entièrement).

* A.V.L. : - les classes difficiles à interpréter globalement se subdivisent en sous-classes ;
- certaines autres se désagrègent totalement.

3) Comparaison des classes :

Nous comparerons en fait les classes ou sous-classes telles qu'elles sont définies dans les récapitulatifs.

a) classes similaires communes :

- logements d'une pièce au centre ville ;
- élément de confort intérieur (ABC) avec maisons luxueuses (AVL) ;
- maisons classiques ;
- terrains ;
- pavillons préfabriqués.

b) réunion de classes :

- transactions autres que ventes (ABC) englobe : grandes propriétés (AVL), offres de location (AVL) et demandes de location et d'achat (AVL).

- environnement du logement (ABC) englobe : petits appartements confortables (AVL) et une partie des appartements moyens (AVL).

- appartements moyens (ABC) englobe : quatre pièces (AVL) et deuxième partie des appartements moyens (AVL).

c) classe dispersée :

Propriétés de campagne (AVL).

4) Conclusion :

La cohésion des classes est plus forte avec la classification basée sur la corrélation. Avec A.V.L., on est parfois obligé de revenir à des sous-classes, mais on ne peut rester à ce niveau car d'autres classes éclateraient totalement. Malgré tout, ces deux classifications sont assez proches l'une de l'autre.

4 - CLASSIFICATION DES ANNONCES :

Le programme de classification existant en bibliothèque [5] ne permet de classifier que trois cent cinquante individus à la fois. L'algorithme utilisé a été celui de la vraisemblance de lien (AVL), celui basé sur la corrélation demandant beaucoup trop de temps calcul. Pour donner un ordre de grandeur, AVL nécessite déjà près de quarante minutes CPU sous Multics.

4-1 Classification avec les 97 attributs existants :

4-1-1 Classification des 350 premières annonces :

L'arbre obtenu comporte deux cent quarante huit niveaux dont quatre vingt trois de significatifs. Etant donné le nombre d'annonces (350), il est ici hors de question de couper l'arbre à un niveau donné et d'énumérer toutes les classes obtenues.

4-1-1-1 Coupure de l'arbre à un bas niveau (47) :

Ce niveau correspond à un maximum local prononcé de la statistique locale. Nous n'y obtenons que vingt cinq classes de plus de trois annonces. En voici quelques-unes :

* à vendre Rennes studio prix 15m
- à vendre studio rue de Saint-Brieuc exposition ouest prix 18m
- à vendre Rennes studio neuf prix 18m
- à vendre Rennes studio 22m
- à vendre près place Ste-Anne studio rénové
- ...

Il est intéressant de remarquer que cette classe correspond exactement à la première classe des deux classifications d'attributs.

* à vendre Gevèze terrain constructible 2000m2 viabilité à proximité
- à vendre Bais terrain 2000m2
- à vendre Gevèze terrain 2000m2, 18m
- à vendre Messac terrain constructible de 2457m2
- à vendre Liffré terrain à bâtir 2050m2 hors lotissement

Comme on s'en aperçoit, ces classes sont significatives, mais de nombreuses annonces restent isolées.

4-1-1-2 Coupure de l'arbre à un niveau élevé :

Dans ce cas, on peut mettre chacune des annonces dans une classe digne de ce nom. Mais ces classes ainsi formées ne sont pas toutes bien homogènes : par exemple le nombre de pièces diffère assez souvent pour les annonces d'une même classe.

4-1-2 Classification des 350 annonces suivantes :

Elle est identique à la précédente ; il est toutefois à noter que les offres de vente sont moins nombreuses au profit de transactions portrant sur les locations.

4-2 Modification des attributs - Nouvelle classification :

4-2-1 Nouvelle liste d'attributs :

Si certaines classes se sont bien dégagées, d'autres en revanche restent confuses. Une des causes nous sembla être le fait que tous les attributs aient la même importance. Prenons par exemple les trois annonces suivantes :

A1: à vendre appt T2, cave, parking, chauffage individuel gaz
A2: à vendre appt T7, cave, parking, chauffage individuel gaz
A3: à vendre appt T7, 130m2, terrasse, garage, à Rennes

L'annonce A2 va très rapidement s'associer avec l'annonce A1 : sept attributs en commun contre trois seulement avec A3.

Dans le même ordre d'idées, d'autres attributs peuvent mener à des associations douteuses :
* attributs d'"état" : beau, récent, libre ...
* attributs "relatifs": prix intéressant, grand, petit ...

Afin d'essayer d'y remédier, nous avons doublé certains attributs (type de l'objet, mode de transaction, nombre de pièces ...) et éliminé certains autres (ceux qui sont susceptibles de conduire à des associations douteuses ainsi que ceux qui apparaissent moins de vingt fois). Nous n'avons alors retenu que soixante dix sept attributs dont la liste figure en annexe.

4-2-2 Classification obtenue :

Nous avons classifier la deuxième tranche de trois cent cinquante annonces. L'arbre produit comporte deux cent onze niveaux pour soixante huit noeuds significatifs. Ici aussi, vu le nombre d'annonces, il n'est pas question de citer toutes les classes obtenues en coupant l'arbre à un niveau donné.

4-2-2-1 Arbre coupé à un bas niveau :

La coupure a été telle que les classes qui en résultent englobent la totalité des trois cent cinquante annonces et soient de cardinal d'au plus dix, un individu isolé pouvant constituer une classe. Il est à noter que l'Algorithme de la Vraisemblance de Lien génére des classes relativement bien équilibrées [1]. Quatre vingt seize classes, toutes très cohérentes ont été isolées ; en voici quelques-unes :

* à vendre Rennes rue de Lorient garage 25300frs
- à vendre garage 45000frs Rennes est
- vends garage quartier sud-est

* à louer quai Lamartine appt T5 de standing
- à louer très beau F5 Rennes
- à louer au centre ville beau 5 pièces, confort, libre de suite, 2500frs par mois.

* étudiant cherche à louer chambre, prix modéré
- cherche à louer studio ou chambre meublée avec cuisine
- cherche à louer petit studio près du centre.

* vends proximité bd J.Cartier maison 4 pièces + cuisine, garage, jardin
- à vendre maison 4 pièces, bd Voltaire, garage et jardin
- à vendre maison rue P.Bourget, T4 sur sous-sol, garage et jardin, prêt possible

* à vendre appartement T2, terrasse, 28m, quartier Bourg-l'Eveque
- Villejean à vendre appt T2 dans petit collectif, 3 étage
- appartement T2 à vendre rue de Brest, bon état, libre à la vente

* jeune couple cherche à louer F3, loyer modéré, Bruz ou environs
- fonctionnaire cherche à louer T3, quartier gare
- jeune couple cherche à louer F3, Rennes ou environ

4-2-2-2 Arbre coupé à un niveau plus élevé :

La coupure a eu lieu à un niveau supérieur à cent cinquante trois de manière à englober dans une classe d'au moins trois éléments chacune des trois cent cinquante annonces. Les quarante six classes obtenues sont (bien sûr) formées de réunions des classes précédentes qui se regroupent généralement bien. Quelques petits accrocs ont cependant lieu comme par exemple la réunion des offres de location de petits logements et de ... maisons luxueuses (ce qui pour ce cas peut s'expliquer par la très faible proportion d'offres de location).

5 - AFFECTATION D'ANNONCES COMPLEMENTAIRES :

5-1 But :

Fournir à un utilisateur ayant composé son annonce une série d'annonces pouvant répondre à ses besoins. Nous avons choisi de proposer au plus (et souvent beaucoup moins) dix petites annonces à l'utilisateur ; il serait en effet maladroit de le submerger de propositions. Le petit système proposé doit être interactif et par conséquent fournir une réponse dans un temps très court.

5-2 Principe de fonctionnement :

Nous supposerons l'existence sur fichier d'un certain nombre d'annonces attendant une réponse (nos trois cent cinquante annonces) partitionnées en classes par la classification automatique (nos quatre vingt seize classes). Dans un souci de rapidité, l'annonce entrée par l'utilisateur n'est comparée qu'à une seule annonce par classe : celle de plus forte dispersion. Le critère de comparaison est l'indice de proximité défini au chapitre 1. La classe dont le représentant a fait le meilleur score est alors présentée à l'utilisateur. Ce système a cependant un point faible : si aucune annonce réellement satisfaisante n'est présente dans la base de données, ce sont les annonces les ... moins lointaines qui sont sorties.

Déroulement du programme

(a) initialisation des petites annonces
(b) initialisation des classes
(c) initialisation des représentants des classes
(d) lecture de l'annonce proposée
(e) analyse et codage de cette annonce
(f) recherche de la classe la plus proche
(g) affichage de cette classe
(h) si une autre annonce est à analyser alors
retour en (d) sinon stop.

5-3 Exemples de fonctionnement :

entrez votre annonce

à vendre garage 25000frs Rennes sud.

annonces complémentaires :

* particulier achète garage secteur Bréquigny

voulez-vous entrer une autre annonce ?

oui

entrez votre annonce

cherche à louer maison de 5 pièces à environ 15km de Rennes

* à louer Pacé maison 5 pièces, garage, jardin, 2000frs/mois
* à louer Pacé maison de 5 pièces
* maison de campagne à louer à 10km de Rennes
* à louer grande maison 5 pièces, salon, séjour, 25km de Rennes, 2500frs/mois

...

à louer F1 centre ville.

* salarié cherche à louer F1 Rennes
* médecin cherche à louer F1 ou F2 Rennes centre

...

achète Rennes nord maison 5 pièces avec prêt.

* à vendre maison T5, Rennes La Bellangeraie, reprise prêt important
* à vendre quartier Patton pavillon T5 avec prêt

...

achète terrain constructible 10km de Rennes environ.

* à vendre à 14km au nord de Rennes, terrain constructible 65000frs
* à vendre Le Rheu, terrain à bâtir 157965frs
* à vendre 15km de Rennes terrain à bâtir 425m2, 90000frs

...

achète appartement F4 de standing, offre 600000frs.

* à vendre appt T4 au centre ville, neuf, terrasse, 633000frs, possibilité de prêt
* à vendre rue de Fougères appt T4 dans petite résidence de grand standig, grand balcon, 59m

...

FIN

5-4 Problèmes non abordés :

Ce petit système se veut avant tout une application aux bons résultats de l'Algorithme de la Vraisemblance de Lien. Aucun des problèmes d'utilisation effective n'a été étudier :

- équité :
  ne pas favoriser une formulation d'annonce par rapport à une autre.

- suppression d'une annonce de la base de données

- ajouts d'annonces isolées dans la base :
  comment les affecter à une classe déjà existante, voire en créer une nouvelle.

- réorganisation totale :
  à partir de combien de modifications (ajouts ou suppressions) réorganiser totalement les classes avec une nouvelle classification par l'algorithme de la vraisemblance de lien.

- détermination automatique des classes d'annonces après une classification.

6 - CONCLUSION :

Nous nous sommes rendus compte que la classification des annonces faite en doublant certains attributs était bien meilleure. Toutefois il reste difficile de voir la différence entre classes regroupant des objets identiques (par exemple les classes d'appartements T4) : nous avons peut-être éliminé trop d'attributs.
Il serait intéressant d'essayer d'améliorer cette méthode. Pour cela, il suffirait d'introduire un vecteur pondération dont chaque composante indiquerait le poids qu'il faut donner à l'attribut correspondant (c'est à dire le nombre de fois qu'il faut appliquer cet attribut). Le calcul des indices de proximité n'en serait pas rallongé pour autant, cela reviendrait à ajouter le poids à la place de un dans les calculs. Ceci, envisageable à moyen terme, demanderait de redéfinir la liste des attributs (si-possible avec des spécialistes des petites annonces) et permettrait certainement d'obtenir une classification beaucoup plus fine.

Dans un avenir plus lointain, nous pourrions abandonner l'attribut de description au profit d'une variable descriptive pouvant prendre en compte une proximité entre ses modalités (par exemple T6 est plus proche de T5 que de T1)

Dans l'immédiat, nous venons de classifier mille annonces à la fois, nous travaillons de manière à obtenir automatiquement les paramètres du petit système à partir des résultats de la classification.

ANNEXE 1 : LISTE DES 97 ATTRIBUTS

Rennes centre	Rennes sud-est
Rennes sud	Rennes sud-ouest
Rennes nord	Rennes est
Rennes ouest	r <= 15 km
15 < r <= 40 km	Rennes divers
région Saint-Malo	autre lieu
1-2 étage	3-4 étage
étage > 4	à vendre
à louer	achète
cherche à louer	jardin
campagne	prix intéressant
prêt	plus frais
libre	collectif
chambre	immeuble
cuisine	meublé
maison	appartement
studio	pavillon
clos	local
terrain	récent
particulier	gaz
neuf	bon - parfait état
à restaurer	restauré
confort	balcon
sous-sol	standing
constructible	beau
aménagé	dépendance
indépendant	plain-pied
calme	grand
petit	en pierres
individuel	garage
cheminée	chauffage
parking	séjour
cave	ascenseur
salle de bain	rez de chaussée
terrasse	cellier
séchoir	grenier
1 pièce	2 pièces
3 pièces	4 pièces
5 pièces	6-7 pièces
plus de 7 pièces	s <= 50 m2
51 - 70 m2	71 - 90 m2
91 - 110 m2	111 - 150 m2
151 - 500 m2	501 - 1000 m2
1001 - 2000 m2	2001 - 3000 m2
s > 3000 m2	p <= 20 m
21 - 30 m	31 - 50 m
51 - 60 m	p > 60 m
1 <= 1000 frs	1001 - 2000 frs
1 > 2000 frs	

ANNEXE 2 : LISTE DES 77 ATTRIBUTS

Rennes centre	Rennes sud-est
Rennes sud	Rennes sud-ouest
Rennes nord	Rennes est
Rennes ouest	r <= 15 km
15 < r <= 40 km	Rennes divers
région Saint-Malo	autre lieu
1 pièce	2 pièces
3 pièces	4 pièces
5 pièces	6-7 pièces
plus de 7 pièces	p <= 20 m
21 - 30 m	31 - 50 m
p > 50 m	s <= 60 m2
61 - 80 m2	81 - 100 m2
101 - 150 m2	151 - 500 m2
501 - 1000 m2	1001 - 2000 m2
s > 2000 m2	à vendre
achète	à louer
cherche à louer	chambre
studio	local
appartement	maison
pavillon	collectif
immeuble	cave
cheminée	chauffage
individuel	gaz
terrain	garage
parking	jardin
grenier	sous-sol
prêt	constructible
terrain seul	terrain seul 2
garage seul	garage seul 2
1 pièce 2	2 pièces 2
3 pièces 2	4 pièces 2
5 pièces 2	6-7 pièces 2
plus de 7 pièces 2	région de Saint-Malo 2
r <= 15 km 2	15 < r <= 40 km 2
à vendre 2	achète 2
à louer 2	cherche à louer 2
chambre 2	studio 2
local 2	

ANNEXE 3 : BIBLIOGRAPHIE

[1] FERNANDO DA COSTA NICOLAU Critérios de análise classificatória hierárquica baseados na função de distribuição - LISBOA 1980.

[2] F. KERLAN rapport de D.E.A. juin 1982.

[3] I.C. LERMAN Classification et analyse ordinale des données Dunod 1981.

[4] I.C. LERMAN Combinatorial analysis in the statistical treatment of behavioral data Quality and quantity, 14(1980) 431-469.

[5] I.C. LERMAN Programme de classification hiérarchique. Rapport I.R.I.S.A. numéro 148 juin 1981.

[6] B. TALLUR Un nouvel algorithme de classification hiérarchique des éléments constitutifs de tableau de contingence basé sur la corrélation. Rapport I.R.I.S.A. numéro 177 juillet 1982.

[7] I.R.I.S.A. - HAVAS cahier des charges novembre 1981.

[8] Projet HAVANE rapport de la convention de recherche I.R.I.S.A. - HAVAS juin 1982.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique