6èmes Journées Francophones
d'Extraction et de Gestion des Connaissances

Lille

4e Atelier
Visualisation et Extraction de Connaissances

le 17 Janvier 2006

François Poulet
ESIEA - Pôle ECD
38, rue des Docteurs Calmette et Guérin
Parc Universitaire de Laval-Changé 53000 Laval

Bénédicte Le Grand
Laboratoire d'Informatique de Paris 6
8, rue du Capitaine Scott
75015 Paris

Avant-propos

Les outils de visualisation contribuent à l'efficacité des processus mis en œuvre en extraction de connaissances en offrant aux utilisateurs des représentations intelligibles et facilitant l'interaction.

La visualisation intervient à différentes étapes de la chaîne de traitement : dans les phases amonts pour appréhender les données et effectuer les premières sélections, lors du processus de fouille, et dans la phase aval pour évaluer les résultats obtenus et les communiquer. Du fait de l'importance croissante accordée au rôle de l'utilisateur en fouille de données, les outils de visualisation sont devenus des composantes majeures des logiciels qui s'utilisent de plus en plus en coopération étroite avec des méthodes automatiques, à la fois en pré et post -traitement.

La fouille visuelle de données ("Visual Data Mining") qui vise à développer des outils interactifs adaptés au traitement des données et des connaissances associées intègre par essence des concepts issus de disciplines diverses : perception visuelle, psychologie cognitive, métaphores de visualisation, visualisation scientifique ou d'information, etc.

Le but de l'atelier Visualisation et Extraction des Connaissances est de fournir un lieu d'échange et de présentation de méthodes nouvelles, d'axes de recherche, de développements dans le domaine de la visualisation en extraction de connaissances et de la fouille visuelle de données.

Les éditions précédentes de l'atelier ont illustré le besoin et l'intérêt grandissant pour la visualisation en extraction de connaissances. Elles ont également permis d'aborder le délicat problème de l'évaluation des méthodes visuelles proposées.

Cette quatrième édition s'inscrit dans la continuité, un tiers des articles s'intéressant à l'évaluation des systèmes de visualisation exploratoire. Les sessions sont organisées en fonction des spécificités des méthodes présentées.

L'atelier Visualisation et Extraction de Connaissances 2006 est composé de quatre sessions :
- Visualisation 3D,
- Interfaces utilisateur – aspects cognitifs,
- Visualisation et analyse de données,
- Visualisation 2D.

Table des matières

Session 1 : Visualisation 3D

Visualisation 3D de paramètres d'apprentissage et de distances

Lionel Martin, Matthieu Exbrayat

Visualisation interactive tridimensionnelle de documents hiérarchiquement structurés

Lila Abes

Visualisation 3D de paramètres d'apprentissage et de distances

Lionel Martin, Matthieu Exbrayat

LIFO, Université d'Orléans
Rue Léonard de Vinci B.P. 6759
45067 ORLEANS Cedex 2 - France
{Lionel.Martin, Matthieu.Exbrayat}@univ-orlens.fr
http://www.univ-orleans.fr/lifo/Members/{martin/exbrayat}

Résumé. Nous présentons dans cet article un outil de visualisation 3D, reposant sur des méthodes usuelles de construction de sous-espaces de représentation (ACP, MDS). L'originalité de notre approche repose sur l'utilisation de ce type de représentation dans le but de déterminer une métrique appropriée dans un cadre d'apprentissage supervisé : nous montrons en effet que l'organisation induite par une mesure de distance (ou de similarité) est une information très riche, comparée par exemple au taux de bonne classification obtenue avec cette métrique.
Cet outil est donc indépendant du type d'objets considérés puisqu'il repose uniquement sur la donnée d'une mesure de distance ou d'une information de même nature. Nous illustrons l'intérêt de cette approche avec deux expérimentations, l'une basée sur des distances entre images, l'autre sur des distances entre molécules décrites en logique du premier ordre.

1 Introduction

Il est admis en apprentissage symbolique ou numérique, que la donnée d'une mesure de distance (ou similarité) appropriée au domaine étudié, constitue un élément clé pour bon nombre de tâches. Traditionnellement, dans un cadre d'apprentissage supervisé, l'adéquation d'une métrique à un type de données est validée par le biais d'un calcul de plus proche voisin appliqué à un jeu de test. La métrique est alors jugée satisfaisante si une majorité d'objets est classée dans la bonne classe, c'est à dire si leur(s) plus proche(s) voisin(s) selon la métrique utilisée appartient (appartiennent) à sa propre classe. Cette méthode permet d'obtenir une information synthétique et numérique, donc facilement exploitable par un logiciel, mais dont l'analyse reste assez limitée. En effet, elle n'offre aucune vue d'ensemble sur les distances entre objets, n'apporte aucune information sur les distances inter classes, ni sur l'existence de plusieurs sous-classes, ni sur les cas litigieux provenant de classes proches ou mêlées et ne distingue donc pas les simples cas limites de véritables outliers.

Dans cette optique nous jugeons pertinent d'introduire l'utilisation d'outils graphiques de visualisation des distances inter objets. Un outil plaçant les objets dans l'espace en respectant le plus fidèlement possible les distances obtenues par le biais de la métrique met en évidence l'organisation induite par la métrique et peut ainsi permettre d'évaluer visuellement la

cohérence des classes (objets proches ou éloignés, classes mélangées ou non) et le repérage des outliers.

Des outils et bibliothèques permettant de telles visualisations existent, reposant notamment sur l'analyse en composantes principales ou le positionnement multidimensionnel. Toutefois, ces outils sont plutôt orientés vers une simple représentation d'un résultat à un instant donné et aucun d'entres eux, à notre connaissance, ne permet d'intégrer ce type de visualisation au sein d'un processus plus global d'apprentissage.

Notre contribution consiste donc tout d'abord à offrir la possibilité d'intégrer la visualisation spatiale dans un cadre plus général, permettant ainsi une interaction directe entre le paramétrage des métriques et la visualisation des distances ainsi calculée. Au delà de ce premier apport, notre logiciel est conçu de façon à pouvoir également paramétrer la visualisation. Il permet enfin de placer des objets de tests dans un espace où sont représentés des objets de classe connue.

2 Vue d'ensemble

L'objectif de l'application est de proposer une représentation 3D de données à partir de sources de données variées : points dans R^3, matrice de distance ou objets décrits dans un espace numérique de dimension éventuellement élevée. Dans tous les cas, le but est de représenter les objets dans le sous-espace de dimension 3 qui offre la meilleure représentation possible, dans le sens où les distances entre objets représentés seront les plus proches possible des distances initialement fournies.

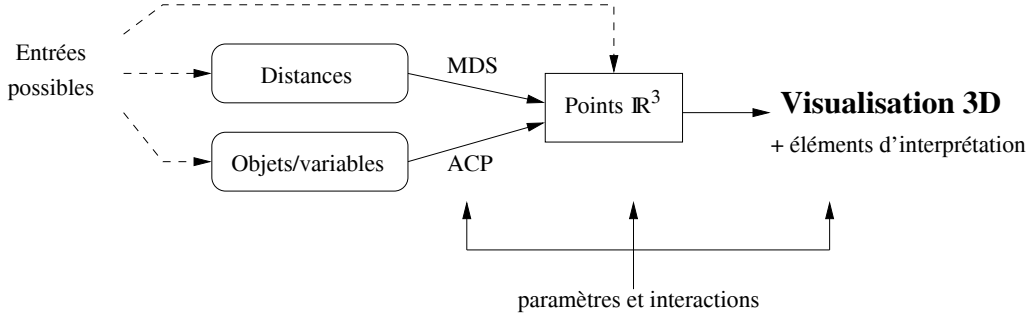


FIG 1 : *Vue d'ensemble de Explorer-3D*

Nous avons choisi une représentation symbolique des objets: chaque objet est représenté par un symbole 3D (sphères, cubes, ...). Par ailleurs, en général les dimensions du sous-espace de représentation sont elles aussi symboliques dans la mesure où il n'est pas nécessairement possible d'en donner une caractérisation.

Dans tous les cas, il est possible de prendre en compte un certain nombre de paramètres et de données spécifiques à un contexte d'apprentissage supervisé:

- il est possible de spécifier un attribut de classe, permettant de représenter les objets d'une même classe avec des caractéristiques visuelles (forme ou couleur) particulières,

- il est également possible de ne prendre en compte que les éléments d'un ensemble d'apprentissage pour déterminer le sous-espace de représentation, puis d'y projeter tous les objets, y compris ceux d'un ensemble test; il suffit pour cela de spécifier un attribut « test » et de choisir la valeur de cet attribut correspondant aux objets de l'ensemble test,
- il est enfin possible de visualiser des informations complémentaires pour chaque objet, affichées en cliquant sur l'objet: il faut pour cela spécifier un attribut dont la valeur (pour chaque objet) correspond au nom du fichier contenant les informations complémentaires. Les types de fichiers visualisables sont pour l'instant des images et du texte.

L'implantation en Java est basée sur Java3D qui permet une interaction très intuitive à la souris: rotation dans les 3 dimensions, zoom, ... L'utilisateur peut également choisir une partie des objets (correspondant à une région de l'espace 3D) et recalculer le meilleur sous-espace de représentation pour ces objets.

2.1 Visualisation d'une matrice de distance

Dans le cas où une matrice de distances est fournie, un positionnement multidimensionnel (MDS) est effectué: pour l'instant, nous nous limitons au cas d'un MDS euclidien, c'est à dire sous l'hypothèse qu'il existe un espace euclidien (de dimension éventuellement élevée) dans lequel il existe un positionnement compatible avec les distances fournies (Kruskal 1978). En plus de la visualisation 3D, plusieurs éléments d'interprétation sont proposés: diagramme de Shepard, contributions, distribution des valeurs propres, ...

Ces éléments d'interprétation fournissent des informations très utiles pour mesurer l'adéquation entre la représentation et les données fournies. Ils permettent également de déterminer si l'hypothèse faite pour le MDS euclidien est satisfaite par la matrice fournie. En effet, la matrice fournie peut ne pas correspondre à une distance, dans le sens par exemple où l'inégalité triangulaire n'est pas satisfaite. Ce cas peut être détecté grâce au diagramme de Shepard; notons cependant que dans cette situation, la représentation 3D peut présenter des distorsions avec la matrice fournie; une implantation de méthodes de positionnement appropriées sera effectuée ultérieurement (Langford et al. 2000, Roweis et al. 2000}).

2.2 Visualisation d'une matrice objets/variables

La détermination d'une mesure de distance (ou de similarité) appropriée peut constituer une étape déterminante dans un processus d'apprentissage supervisé. De nombreux travaux proposent une modélisation de ce type de mesure, soit basée sur une mesure entre les descriptions des objets (Bisson 1992, Emde et al. 1996), soit basée sur une somme (éventuellement pondérée) d'attributs numériques (éventuellement 0 ou 1). Ces attributs peuvent être:

- soit issus des attributs disponibles sur les objets (dans le cas attribut-valeur) (Aha 1989, Giraud-Carrier et al. 1995),
- soit synthétisés (ou appris) dans le cas d'objets décrits en logique du premier ordre (Sebag 1997, Sebag et al. 1994, Domingos 1995, Martin et al. 2001).

Dans ces cas, on dispose d'une représentation objets/variables qui se prête particulièrement bien à l'analyse en composantes principales (Lebart et al. 2000), c'est pourquoi nous proposons de prendre en charge ce type de données. Dans ce cas aussi, Explorer-3D permet de visualiser les objets en 3D, chaque dimension correspondant aux composantes principales choisies par l'utilisateur. Il est possible d'effectuer une analyse en données non réduites afin de prendre en compte une éventuelle pondération des variables. Enfin, les éléments d'interprétation usuels en ACP (contributions, cos carrés, ...) peuvent être analysés.

3 Expérimentations

Comme nous l'avons précisé, Explorer-3D peut être utilisé sur n'importe quel type d'objet, à condition de donner ou de définir une mesure de distance (ou de dissimilarité) sur ces objets. Pour illustrer l'intérêt de Explorer-3D, nous proposons deux applications sur des données très différentes :

- la première présente l'organisation associée à des mesures de distances entre images médiévales, ces distances étant uniquement définies à partir des couleurs communes aux images;
- la seconde permet de visualiser l'organisation de molécules décrites en logique du premier ordre: ce jeu de données sur le caractère mutagène des molécules est très utilisé en Programmation Logique Inductive (ILP) (Srinivasan et al. 1994).

3.1 Classification d'images à partir des couleurs

Une des application que nous avons mis en place consiste à évaluer des distances ou dissimilarités entre des enluminures médiévales. Ces mesures utilisent les couleurs des enluminures afin d'effectuer un classement par époque et région d'origine. En effet, les pigments utilisés varient au cours du temps et suivant les régions. Nous avons étudié deux mesures reposant sur des histogrammes tridimensionnels de couleurs, avec l'espace colorimétrique RVB.

Nos tests reposent sur des images d'enluminures numérisées à l'aide de scanners calibrés, fournies par l'Institut de Recherche et d'Histoire des Textes (IRHT).

3.1.1 Mesure reposant sur la proportion de couleurs communes

La première mesure, appelée dc_1, repose sur la proportion de couleurs communes. Soit un histogramme P_a de présence des couleurs dans l'image I_a, nous avons P_a[i,j,k]=1 si l'enluminure comporte au moins un pixel de teinte (R=i,V=j,B=k), et P_a[i,j,k]=0 sinon. Posons *nb1* et *nb2* le nombre de couleurs présentes respectivement dans P_1 et P_2, et *nbc* le nombre de couleurs communes à ces deux histogrammes. La mesure dc_1 se calcule comme suit :

$$dc_1(P_1, P_2) = \frac{nbc1 + nbc2 - nbc}{nbc} - 1 = \frac{nbc1 + nbc2}{nbc} - 2$$

Dans le cas où les histogrammes ne comportent pas de couleurs communes, nous posons une distance maximum $dc_1(P_1,P_2)$=card(P).

dc_1 ne constitue pas une distance à proprement parler, mais plutôt une mesure de dissimilarité. En effet, elle ne vérifie pas la propriété d'inégalité triangulaire (pour trois histogrammes A,B,C, dc_1(A,B)+dc_1(B,C) ≥ dc_1(A,C) n'est pas toujours vérifié).

3.1.2 Mesure reposant sur les fréquences de couleurs

La seconde mesure, appelée dc_2, dérivée de (Swain et al. 1991), repose sur les fréquences de couleurs. Soit un histogramme F_a de fréquence des couleurs dans l'image I_a, nous avons F_a[i,j,k]=nbp[i,j,k]/NBP, avec nbp[i,j,k] le nombre de pixels de teinte (R=i,V=j,B=k), et NBP le nombre total de pixels dans l'image. Cette expression permet de traiter des images de tailles différentes. Nous avons alors :

$$dc_2(F_1, F_2) = \sum_{i,j,k} \left| F_1[i, j, k] - F_2[i, j, k] \right|$$

dc_2 constitue une distance (il s'agit en fait d'une distance de Manhattan).

3.1.3 Visualisation des distances entre histogrammes

Le test présenté ici repose sur un jeu de 166 enluminures provenant majoritairement de France et datées du 12ème au 15ème siècle. Les images sont extraites de 10 manuscrits différents, certains manuscrits provenant de la même région et de la même époque.

Les figures 2 et 3 présentent respectivement la visualisation des mesures obtenues à l'aide de dc_1 et de dc_2. Lors de nos manipulations, nous utilisons essentiellement les couleurs pour distinguer les différentes classes. Toutefois, afin d'aider l'interprétation des copies d'écran, nous avons utilisé des formes (cube, cône, cylindre, sphère) pour mieux distinguer les différentes sources (époques). Nous pouvons noter que pour dc_2, les objets sont bien répartis dans l'espace et forment des groupes homogènes suivant les dates (et également suivant les lieux de production). Il est à noter que certains objets mal placés correspondent en réalité, après vérification auprès de l'IRHT, à des enluminures ajoutées tardivement dans les manuscrits, et dont l'origine diffère donc du reste du document. Avec dc_1, le placement est plus difficile à interpréter. Nous estimons que cela provient du fait que dc_1 ne constitue pas une véritable mesure de distance, ce qui est en contradiction avec l'hypothèse prise pour notre implantation du positionnement multidimensionnel. Toutefois, les groupes restent homogènes.

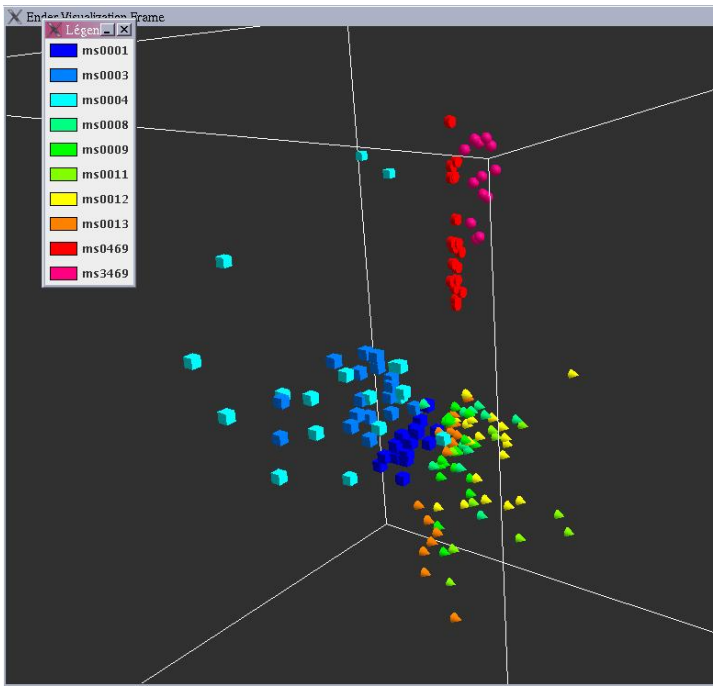


Fig 2 : *Visualisation des distances entre enluminures avec* dc_1

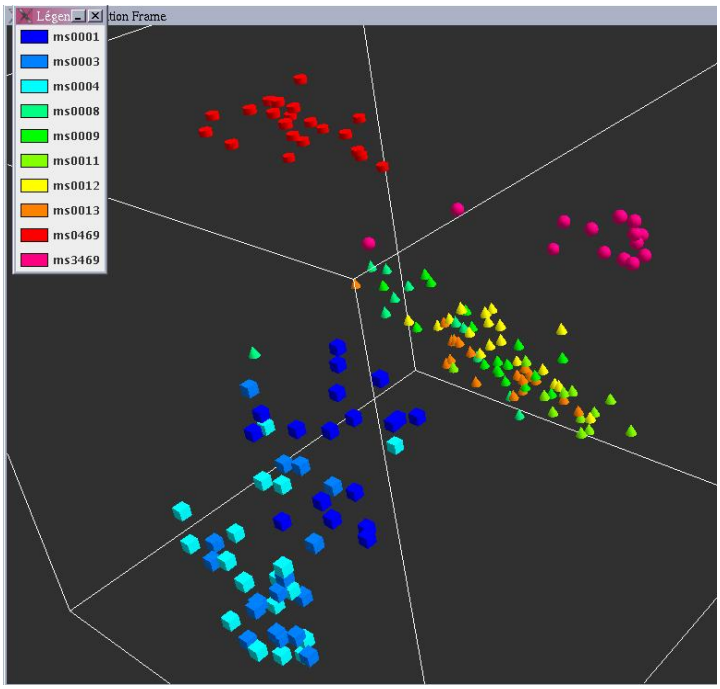


Fig 3 : *Visualisation des distances entre enluminures avec* dc_2

3.2 Mutagenesis

Mutagenesis (Srinivasan et al. 1994) est une base de données décrivant la structure de 230 molécules. Cette description est réalisée en premier ordre avec des prédicats concernant la nature des atomes de chaque molécule, les liaisons entre ces atomes ainsi que des informations complémentaires. Parmi ces informations figure un attribut indiquant le caractère mutagène ou non de la molécule, qui constitue l'attribut de classe.

Ces dernières années, de nombreux articles en ILP ont utilisé ce jeu de données pour tester les approches proposées. Les résultats obtenus sont finalement assez proches en terme de taux de classification correcte (environ 90 %).

Comme nous l'avons mentionné, les taux de bonne classification constituent une indication importante mais finalement, ils résultent d'une phase de mise au point au cours de laquelle un certain nombre de paramètres ont été fixés. La classification d'une molécule à l'une ou l'autre des classes peut parfois être qualifiée d'incertaine, par exemple si, dans une approche de type plus proche voisin, les 2 plus proches voisins d'une molécule sont de classes différentes, et à distance comparable. Nous proposons ici une information beaucoup plus riche du résultat de classification grâce à la visualisation.

Nous considérons ici des distances basées des langages (Martin et al. 2001): le langage choisi est associé à un nombre donné de termes représentant des propriétés (associées aux variables de la matrice objets/variables), avec une valeur 1 si la molécule vérifie la propriété et la valeur 0 sinon. Si nous ne considérons dans un premier temps que des propriétés assez discriminantes, nous obtenons une distance à partir de laquelle la représentation induite est présentée figure 4.

Cette projection en 2D montre au mieux la répartition dans le plan; chaque classe est associée à une couleur (claire ou sombre). On peut ainsi observer une région dans laquelle se concentrent beaucoup de molécules des 2 classes, dans cette région, les molécules sont donc très proches (similaires) les unes des autres.

Si nous considérons maintenant une distance associée à un ensemble de termes correspondant à des propriétés syntaxiquement simples (et non nécessairement discriminante), nous obtenons la répartition de la figure 5.

Bien que les 2 mesures donnent des résultats comparables en terme de taux de bonne classification, l'organisation obtenue dans le deuxième cas semble plus riche puisqu'on peut observer:

- des groupes ``purs" de chaque classe, ne comportant que des objets d'une seule classe; pour ces groupes, la classification par plus proche voisin sera fiable, quelque soit la taille du voisinage,
- un groupe contenant des objets des 2 classes; même si on peut distinguer une frontière entre les 2 classes, des erreurs de classification dans ce groupe auront vraisemblablement lieu, il faudrait faire une analyse plus fine de ce groupe,
- enfin quelques objets (clairs) au milieu d'un groupe d'objets de l'autre classe; ce cas correspond à des objets que l'on peut qualifier d'atypiques.

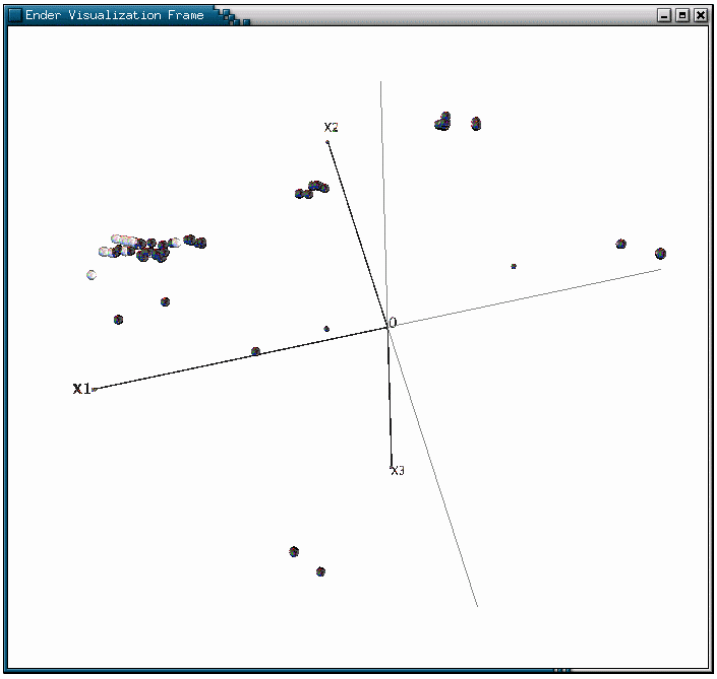


Fig 4 : Mutagenesis: organisation induite par une distance

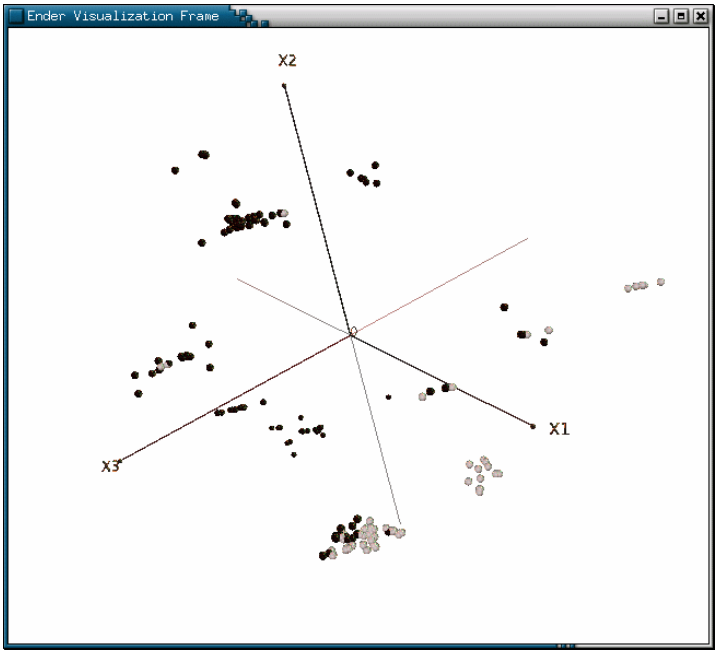


Fig 5 : Mutagenesis: organisation induite par une autre distance

Ces deux figures indiquent l'organisation générale des objets; il faut garder à l'esprit que la projection engendre une perte dans les distances représentées. Dans les deux cas on pourrait faire une étude plus fine en ne représentant par exemple que les objets appartenant à des groupes « non purs », la représentation serait alors plus fidèle.

3.3 Limitations

L'approche présentée ici est confrontée à deux types de limitations lorsque le nombre d'objets est élevé :

- la première limitation concerne le nombre d'objets représentables : chaque objet est représenté par une forme 3D, elle même représentée par un certain nombre de triangles. Il existe nécessairement une limite au nombre d'objets représentables, liée aux capacités mémoire ; par ailleurs, un grand nombre d'objets entraîne une densité élevée et rend l'interprétation plus difficile. Nous avons fait des tests jusqu'à 5000 objets ; cette valeur est sans doute assez proche de la limite aussi bien du nombre d'objets représentables qu'interprétables,
- la seconde limitation est liée au calcul du sous-espace de représentation. Dans le cas d'objets représentés dans un formalisme attribut/valeur, il est nécessaire pour l'ACP de diagonaliser une matrice nxn ou pxp, n étant le nombre d'objets, p le nombre d'attributs (on diagonalise la plus petite matrice). Dans le cas où l'on dispose d'une matrice de similarité, le MDS nécessite la diagonalisation d'un matrice nxn. Nous utilisons la librairie Jama, qui contient des fonctions optimisées pour la diagonalisation (en particulier pour les matrices définies positives, ce qui est notre cas en général). Dans nos expérimentations, les matrices diagonalisées atteignaient une dimension maximum de 250x250, et nécessitaient environ 5 secondes de calcul. Il est sans doute possible de diagonaliser des matrices de quelques milliers de lignes/colonnes.

Si la taille des données dépasse ces limitations, dans les deux cas il sera nécessaire d'échantillonner : cela ne devrait pas trop pénaliser l'observation de l'organisation générale en régions plus denses (il sera alors possible d'étudier ces régions en représentant tous les objets qu'elle contient). De même en cas d'échantillonnage pour le calcul du sous-espace de représentation, le sous-espace obtenu sera une approximation du meilleur sous-espace, l'erreur pouvant alors être bornée (en probabilité).

4 Conclusion et perspectives

Les travaux présentés dans cet article donnent un éclairage nouveau sur les distances utilisées en apprentissage supervisé. En permettant de visualiser l'organisation des objets induite par une distance donnée, notre application offre des possibilités nouvelles dans la mise au point de distances, elle permet également à l'utilisateur d'obtenir des informations sur les données traitées, de repérer des groupes homogènes, des outliers et de classer de nouvelles observations.

Notre objectif est de fournir un outil interactif pour la mise au point de distances et leur visualisation sur un jeu de données. Pour l'instant, l'interaction est essentiellement axée sur le paramétrage des méthodes de réduction de dimensions, sur la visualisation et les choix des paramètres d'affichage. Sur ce point, il reste à implanter des méthodes de positionnement multidimensionnel dans le cas où la « matrice de distance » fournie n'est pas une matrice de distance à proprement parler.

Par ailleurs, notre formalisme de distance basée sur un langage est très général et permet de représenter la plupart des distances utilisées en apprentissage puisqu'il peut être largement paramétré et qu'il peut s'appliquer à n'importe quel type de données. Nous souhaitons étendre l'interaction dans ce sens, permettant ainsi à l'utilisateur de visualiser l'influence de chaque paramètre sur l'organisation induite. A terme, nous envisageons un réglage partiellement automatisé des paramètres du langage (donc de la distance) à travers une interaction directe sur la représentation 3D.

Références

Aha D. (1989), Incremental, instance based-learning of independent and graded concept descriptions, Proceedings of Sixth International Machine Learning Workshop (ML89), pp. 387-391, 1989.

Bisson G. (1992), Learning in FOL with a similarity measure, Proceedings of 11th National Conf. on Artificial Intelligence (AAAI), pp. 82-87, AAAI Press, 1992.

Domingos P. (1995), Rule Induction and Instance-Based Learning: A Unified Approach, Proceedings of Fourteenth International Joint Conference on Artificial Intelligence (IJCAI'95), pp. 1226-1232, Morgan & Kaufmann, 1995.

Emde W. et Wettschereck D. (1996), Relational Instance-Based Learning, Proceedings of 13th Int. Conf. on Machine Learning (ICML'96), pp. 122-130, Morgan & Kaufmann, 1996.

Giraud-Carrier C. et Martinez T. (1995), An Efficient Metric for Heterogeneous Inductive Learning Applications in the Attribute-Value Language, Proceedings of GWIC'94,

Vol. 1, pp. 341-350, Kluwer Academic Publishers, 1995.

Kruskal J. (1978), Multidimentional Scaling, Sage Publ., 1978.

Lebart L., Morineau A. et Piron M. (2000), Statistique exploratoire multidimensionnelle, Dunod, 2000.

Langford J.C., Tenenbaum J.B., Bernstein M. et De Silva V. (2000), Graph Approximations to Geodesics on Embedded Manifolds, 2000.

Martin L. et Moal F. (2001), A Language-Based Similarity Measure, Proceedings of 12th European Conference on Machine Learning (ECML), LNAI, Vol. 2167, pp. 336-347, Springer Verlag, 2001.

Roweis S.T. et Saul L.K. (2000), Nonlinear Dimensionality Reduction by Locally Linear Embedding, Science, Vol. 290, pp. 2323-2326, 2000.

Sebag M. et Schoenauer M. (1994), A Rule-based Similarity Measure, *dans* Topics in Case-Based Reasonning, LNAI, Vol. 837, pp. 119-130, Springer-Verlag, 1994.

Sebag M. (1997), Distance Induction in First Order Logic, Proceedings of ILP'97, pp. 264-272, Springer-Verlag ,1997.

Srinivasan A., Muggleton S., King R.D. et Sternberg M.J.E. (1994), Mutagenesis: ILP experiments in a non-determinate biological domain, Proceedings of the 4th International

Workshop on Inductive Logic Programming, pp. 217-232, Gesellschaft für Mathematik und Datenverarbeitung MBH, 1994.

Swain M.J. et Ballard D.H. (1991), Color indexing, International Journal of Computer Vision, 7(1), pp. 11-32, 1991.

Summary

In this paper we introduce a 3D visualization tool, based on current methods for the construction of representation subspaces (PCA, MDS). One of the most innovative aspects of this approach resides in the fact that this spatial distribution is used as a supervised learning tool, which helps to determine accurate distance parameters and metrics. We show that the spatial organization that can be deduced from distances (or similarities) between objects is a powerful information, compared to more classical evaluation metrics, such as good classification rates.

This tool is independent of the kind of object we study, as it is only based on a distance measurement (or some equivalent data). The interest of this approach is highlighted through two experiments, the first one based on a distance between pictures, and the second one based on distances between molecules that are described in first order logic.

Visualisation interactive tridimensionnelle de documents hiérarchiquement structurés

Lila Abes*
Département communication homme/machine, groupe Architecture et Modèles pour l'Interaction (AMI) ;LIMSI-CNRS, Bat. 508, BP 133, 91403 Orsay, France
*lila.abes@limsi.fr

Résumé. L`article suivant est concerné par deux domaines principaux de recherches : Le Text Mining et la visualisation Interactive de l'information appliquée aux documents hiérarchiquement structurés. Les principales approches existantes tendent à faire correspondre une hiérarchie de documents à un espace de visualisation en utilisant des arrangements graphiques où le critère de performance est l'utilisation de l'espace ou optimisation du rendu visuel. Ici nous proposons une nouvelle approche qui est concernée par la sémantique des distances originales entre les documents aussi bien que par l'optimisation de l'utilisation de l'espace. L` algorithme procède de manière ascendante en calculant a chaque étape des disques reconfigurables correspondants a des paliers (ou noeuds) de la hiérarchie, et en opérant une minimisation interne aux noeuds de même classe et maximisation externe aux noeuds de même palier[1]. Un fichier XML d'entrée-sortie est employé comme format pivot de communication avec d'autres composants : l'outil de classification et la plateforme de visualisation. Des manipulations interactives sont définies dans un cadre spécifique (voir le modèle de Chi&Riedl, FIG.3).

1 Introduction

Les difficultés liées à la gestion de grands ensembles de données semble être une problématique assez récurrente et bien suffisamment érodée. En fait, la littérature fleurit au sujet des méthodes et des techniques concernées par la classification de données et la visualisation de données, les deux considérés séparément. Ce travail peut donc sembler être une énième tentative dans ce sens. En fait, traiter l'information en vue d`une analyse ne relève pas moins d'une bonne structuration que d`une bonne (re)présentation des données à explorer et nous estimons que les deux préoccupations doivent être fusionnées en une seule, tel est donc l`objet du résumé des travaux présentés ici.

Deux approches existent donc lorsqu`il s`agit de visualiser des données hiérarchisées : une première approche s`appuie sur la projection de l`arborescence dans un espace de visualisation de sorte que les propriétés de ce dernier, exprimées en termes d'expressivité et efficacité soient optimisées (Beshers&Feiner, 1997) ; une deuxième manière de faire postule de maintenir les distances entre les données – après projection dans un espace de représentation – aussi proches que possible des distances avant projection (F. Jourdan, G. Melançon, C. Douy et A. Gasne, 2005). Notre but est de formuler et mettre en application une nouvelle approche où la spatialisation des représentants graphiques est non seulement une question d'utilisation de l'espace mais aussi de minimisation de la distorsion des distances originales après spatialisation.

2 Etude comparative et fondements théoriques

Comme mentionné ci-dessus, deux directions pourraient être entreprises pour la définition d`une stratégie de spatialisation, guidé par deux critères :

[1] Voir Fig.1 pour le lexique (définition de noeud et palier).

a. Critère basé sur une meilleure utilisation de l`espace: Les noeuds associés aux classes à projeter sont placés dans un espace bidimensionnel, de sorte que tous les noeuds soient montrés ou puissent être potentiellement montrés (l'espace de visualisation est alors expressif); tandis que tous les noeuds peuvent être accessibles – même cachés - dans le cas d`un environnement tridimensionnel (l'espace de visualisation est alors efficace). L`inconvénient principal est dans ce cas-ci l`insignifiance des distances intérieures entre les objets montrés dans l'espace de projection donné.

b. Critère de divergence minimale des distances initiales entre objets après projection: la graduation multidimensionnelle (MDS) est basée sur une fonction d'optimisation où un ensemble d'objets distingués par des estimations de dissimilitudes est rapporté à une matrice des distances géométriques représentatives (Kruscal, 1979). Ainsi, une fonction de la somme des erreurs, f, donne une approximation de la marge â la solution optimale (f ˜ 0), qui est généralement efficace ; mais une objection peut être émise a savoir le manque d'efficacité dans un environnement interactif dû à la lenteur de l`algorithme MDS et des difficultés liées au cas de taxonomies ou de données non numériques (quelques travaux récents dans ce domaine peuvent être mentionnés comme les algorithmes MDS hybride (Jourdan&Melançon, 2005).

Ce qui nous amène â l`énoncé de notre algorithme, mais commençons d`abord par poser les fondements théoriques ; Soit:

$X = \{X_i / i=1..n\}$ un ensemble de données a classifier
$M= \{di*j (X_i , X_j), i=1..n, j=1..n\}$ la matrice de dissimilitudes codant les degrés de dissimilitudes entre chaque paire de données

$U = \{ U P_i / P_i = U \{ U \{c_j, U\}\ \}, i=1..l, j=1..m\}$ tel que :
P: partitions de classes délivrées par un outil de classification hiérarchisant;
i : un niveau de la hiérarchie
l: nombre de niveaux,
m: nombre de classes dans un niveau i
U: les valeurs ultra métriques, chaque valeur de dissimilitude est calculées en utilisant une mesure d'estimation de distance pour chaque paire de classes,
Alors, nous avons: U = T (d);

L'étape de visualisation vise à calculer les coordonnées des classes c_i dans la matrice U':
$U' = (U, C (X , Y, Z))$
$= \{ U P_i / P_i = U \{ U \{c_j (x_i , y_i , z_i) , U' \}\} \}$
(x_i , y_i , z_i) : coordonnées de c_i dans l`espace de visualisation 3D
U' : nouvelles dissimilitudes induites par U'
Ainsi, on obtient: U' = T' (U)

- Dans le cas (a), la visualisation est fonction de l`optimisation de l`utilisation de l`espace, le Système optimisé est alors:

$U – U = \{ U P_i / P_i = U \{ U \{c_j\}\ \}, i=1..l, j=1..m\}$

Afin de générer U', seul les liens entre classes sont pris en considération. T est alors insignifiant et la déformation des distances originales n`est pas équivoque.

- Dans le Cas (b), la visualisation est fonction de U et la valeur de T peut être estimée, l'opération de projection vise à calculer une autre ultra métrique (défini comme mesure spatiale) T'de sorte que T'- T est réduit au minimum.
Les algorithmes MDS se rapprochent de cette optique, mais à moins d'une distribution de MDS locaux opérant sur chaque niveau, il reste peu convenable pour des hiérarchies et reste trop lent en termes de temps d'exécution.

A ce propos, nous proposons – dans la section suivante - une formalisation d'un algorithme de spatialisation basée sur un calcul optimisateur sans avoir recours a l`échantillonnage multidimensionnel. Nous définissons les entrées, le modèle de table d'entrée-sortie pour le stockage de données en cours de calcul ajusté sur le modèle de table de Card (Card, McKinlay, Schneiderman, 2000) ainsi qu`une description sommaire de son fonctionnement.

3 Description générale de l`algorithme de spatialisation et format des entrées/sorties

Etant donné une hiérarchie de documents issue d`un résultat de classification, le but est alors de projeter chaque noeud de la hiérarchie – représentant une classe – a des noeuds graphiques: en partitionnant l`écran de visualisation (Tree-Map, (Shneiderman, 1992)) accentuant l'expressivité, objets reconfigurables tels les RDT (Jeong&Pang, 1998)) ou zones recouvrantes (galaxies, (Keinraich&Sabol, 2003)) augmentant l'efficacité; ou tout simplement par des points graphiques dont les distances deux a deux sont conformes aux dissimilitudes originales. Dans notre cas, ladite projection est matérialisée par un calcul des coordonnées de représentants graphiques (en l`occurrence ici des disques) assorties d'une table d'entrée-sortie FIG-1], utilisé comme tableau noir par notre algorithme.

Soient Di un disque virtuel matérialisant un niveau de la hiérarchie (colonne de la table); et Ci un disque réel matérialisant une classe:

Attributs de D_i	Attributs de C
. H_i : hauteur de D_i (coordonnée z) . $Ø_i$:diamètre du disque relatif a H_i . $Cord_i$: coordinnees du centre du disque D_i en 3D . N_i : nombre de noeuds fils . Ref_i : reference sur la liste lineaire ordonnée des neouds visibles sur le disque virtuel	- $Char_i$: description de la classe C - $Cordc_i$: coordonnés du centre du disque représentant C en espace 3D . $Øc_i$: diamètre du disque représentant C_i - Tab_i : table des references pointant tous les noeuds fils sur la longueur de la hiérarchie - $Cref_i$: reference sur le disque représentant un noeud frère (classe de même niveau)

Dans la Fig.1, H_i , N_i , $Char_i$, Tab_i , $Cord_i$, $Cref_i$ sont les sorties calculées à partir du fichier XML issu de l'outil de classification (résultat de classification arborescent) ; $Ø_i$, $Cordc_i$, $Ø_{ci}$ sont alors les valeurs à calculer et à mettre à jour itérativement dans la table précédente. En effet, chaque niveau de la table correspond à un palier de la hiérarchie, pour chaque noeud (FIG.1) il s`agira de faire correspondre un disque (FIG.2) tout en respectant les distances initiales entre les noeuds. Pour les disques de même classe, la distance est minimisée pour les garder les plus proches possibles (distance est dans ce cas la plus petite distance reliant chaque paire de classes dans la matrice ultrametrique, modélisée dans la table des

entrées/sorties) et est calculée de manière empirique ; tandis que pour les disques de même palier mais de classes différentes, la distance est cette fois-ci maximisée par l`opération inverse, à savoir, recherche de la plus grande distance reliant les centres de tous les disques auparavant minimisés.

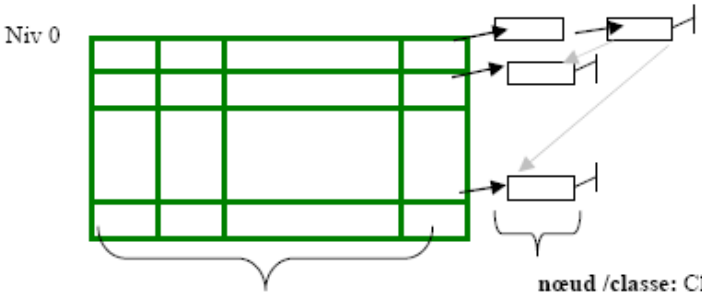


FIG. 1 – Table des entrées/sorties [2]

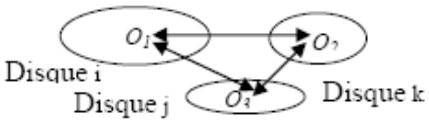


FIG.2 – Système de disque virtuel, représentant des classes par des disques dans l`environnement de visualisation

4 Interactivité

Des opérations interactives sont ici requises, spécialement en espace 3D. Nous proposons deux scénarios principaux : filtrage/exploration de données de la hiérarchie, ajout de nouvelles données. Nous suggérons d'employer un arbre de décision (précédemment entraîné sur l'ensemble original de données) pour filtrer et enrichir la hiérarchie originale, où chaque opération actionnée à un niveau abstrait du modèle de Chi&Riedl est transformée en une opération liée au niveau abstrait immédiatement supérieur :

- Filtrer les données:

1. ***Niveau Vue:*** l'utilisateur présente une requête des données a filtrer (rechercher) sous forme d`un vecteur des intervalles d valeurs des attributs. **2.** ***niveau Abstraction Analytique:***la demande est soumise à l'arbre (l'arbre de décision, associé au niveau d`abstraction analytique). La fonction de décision examine - largeur d'abord - chaque noeud d'arbre et dérive un ensemble d'arbres secondaires de classes qui se rapprochent de la demande donnée. **3.** ***Niveau transformation du Mapping:*** en utilisant l'arbre secondaire, une hiérarchie secondaire est extraite à partir de la table d'entrée-sortie. **4.** ***Niveau transformation Visuelle:*** la visualisation est alors arrangée à la vue correspondante (en employant les coordonnées des classes extraites de la table d`entrées/sorties).

- Ajout de nouvelles données (nouveaux documents):

1. ***Niveau vue:*** l'utilisateur présente de nouveaux (ou ensemble de nouveaux) éléments assortis d'attributs valués. **2.** ***Niveau Abstraction Analytique:*** l'arbre de décision calcule les classes correspondantes aux valeurs des attributs fournis, ou crée les classes correspondantes. **3.** ***Niveau transformation du Mapping:*** les coordonnées des nouvelles classes sont calculées (seuls les niveaux supérieurs sont nécessaires pour le reclassement). **4.** ***Niveau transformation Visuelle:*** l'utilisateur peut fixer graphiquement a l`aide d`objets visuelles et des coordonnées calculées dans le secteur (zones de l`écran) indiqué.

[2] Adapté du modèle défini par Card (Card, Mckinlay Schneiderman, 2000).

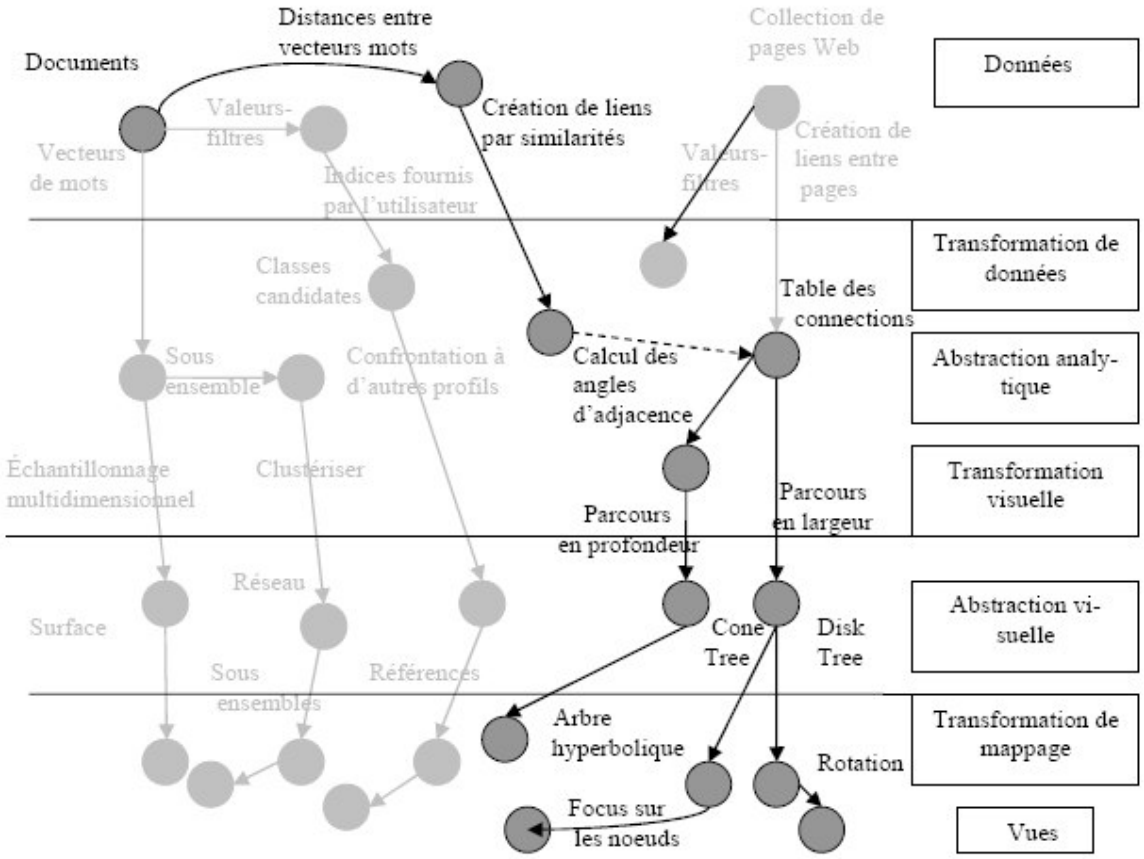


FIG.3- framework d`interactivité basé sur le modèle de Chi&Riedl model [3]

5 Conclusion et perspectives

Dans cet article, nous avons présenté une nouvelle méthode de spatialisation de hiérarchies et nos motivations pour un tel processus. Le paradigme est soutenu par un postulat : une bonne structure représentative graphique pour un arbre de données est celle qui déforme le moins les similitudes originales entre les unités de données (similitudes qui sont donnés par l'outil de classification); ou en d'autres termes : deux individus (ou classes) qui sont proches l'une de l'autre le restent aussi possible que peut après projection dans l`espace de visualisation ; et ceci doit être évident (et vérifiable par une mesure de calcul de divergence) dans l'environnement de visualisation.

A cet effet, quelques expérimentations sont en cours sur un ensemble de données (documents recueillis sur le Web concernant des corporations professionnelles). Une mesure de qualité - liée à T ', qui évalue la déformation maximale induite après projection - est proposée afin de comparer aux algorithmes similaires. Ceci dans une première étape, une deuxième étape est nécessaire pour la visualisation appropriée ; ainsi, certaines opérations interactives telles que l`interrogation/exploration de la hiérarchie, l'extraction de données et l'ajout de nouveaux éléments ont été modélisées sous forme de diagrammes de scénarios où les opérations interactives ont lieu entre l'utilisateur et l'environnement, chaque opération représente une transformation d'un ensemble d'opérations des niveaux abstraits supérieurs du modèle adapté de Chi&Riedl.

[3] Le sous modèle qui nous intéresse est celui mis en exergue.

References

C. Beshers and S. Feiner (1997). Generating efficient Virtual Worlds for Visualization Using Partial Evaluation and Dynamic Compilation. PEPM 1997, P. 107-115.

S. K. Card, J. D. Mackinlay, B. Schneiderman, (1999) .Readings in information visualisation, using vision to think. Morgan Kaufmann publisher.

E. Diday, (1984). Une représentation visuelle des classes empietantes : les pyramides. Rapport de recherche, Inria, France.

E. Diday, (2002). An introduction to Symbolic Data Analysis and SODAS software. J. S. D. A., International E-Journal.

C. Jacquemin, H. Folch and S. Nugier, (2005). Exploration d'analyse de données textuelles et navigation contrôlée dans OCEAN . Actes IHM'05, Toulouse, France.

C-S. Jeong et A. Pang, (1998). Reconfigurable disc trees for visualizing large hiearchical information space. Symposium on Information Visualization, pages 19-25. IEEE.

F. Jourdan, G. Melançon, C. Douy and A. Gasne, (2005). Une approche MDS hybride pour l'exploration visuelle interactive. IHM 2005, Toulouse.

B. Kruskal and M. Wish, (1979). Multidimensionnal scaling. Jin Sage publications, Newbury Park CA.

B. Shneiderman, (1992). Tree visualisation with tree maps: a 2D space filling approach. ACM Transactions on graphics, vol. 11.

Session 2 : Interfaces utilisateur – aspects cognitifs

Évaluation des Interfaces Utilisateur d'Information

Nicolas Bonnel, Max Chevalier

Visualisation de navigation sur interface graphique pour l'analyse cognitive de parcours

Marc Damez, Stephan Renaud

Vers une méthodologie rigoureuse de conception des langages graphiques s'appuyant sur les sciences cognitives

Jean-Baptiste Lamy, Catherine Duclos, Vincent Rialle, Alain Venot

Évaluation des Interfaces Utilisateur d'Information

Nicolas Bonnel*, Max Chevalier**,***

*IRISA / Université de Rennes 1, Campus universitaire de Baulieu, 35042 Rennes Cedex
nicolas.bonnel@irisa.fr
http://www.irisa.fr/texmex/Nicolas.Bonnel

**IRIT / Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex
Max.Chevalier@irit.fr
http://www.irit.fr/~Max.Chevalier

***LGC / IUT Rangueil, 115 route de Narbonne, 31062 Toulouse Cedex

Résumé. Tout processus de recherche d'information (RI) n'a de sens aux yeux des utilisateurs qu'à travers l'ultime étape qui consiste à visualiser les résultats. L'importance que prend la visualisation (ou restitution) des résultats est à l'origine des nombreuses propositions d'interfaces, qu'elles soient textuelles, 2D ou 3D. Si des évaluations de certaines de ces interfaces ont été proposées, aucune comparaison n'a réellement été réalisée dans ce contexte faute de contraintes sur ces interfaces et de critères de comparaison relatifs à la tâche de RI. Dans cet article, nous proposons d'introduire des pistes d'évaluation afin d'aboutir à un cadre expérimental permettant l'évaluation et ainsi la comparaison des ***Interfaces Utilisateur d'Information***.

Mots-clés. Interface Utilisateur d'Information, interface de restitution de résultats de recherche, systèmes de recherche d'information, critères d'évaluation et de comparaison, évaluation « en contexte ».

1 Introduction

Pour rechercher des informations sur le Web, une grande majorité des utilisateurs utilise des moteurs de recherche. Ces outils proposent un frontal grâce auquel l'utilisateur peut spécifier sa requête et en visualiser les résultats. On constate alors un décalage entre le nom de ces outils et leur fonction du point de vue des utilisateurs. C'est pourquoi nous parlerons désormais dans cet article d'**Interfaces Utilisateur d'Information (ou IUI)**. Ce terme a été proposé par Franck Poisson, ex-Directeur Général et fondateur du bureau France de GOOGLE.

L'évaluation des IUI est une tâche particulièrement difficile. Jusqu'ici, la méthode la plus courante est la réalisation d'une étude (ou test) utilisateur. Celle-ci ne nous semble pas suffisante et nous proposons dans les sections suivantes des pistes d'évaluation des IUI. Nous présentons tout d'abord les principales limites des évaluations faites actuellement qui sont

certes très instructives mais qui ne nous permettent pas d'évaluer l'impact qu'a l'IUI sur la tâche accomplie par l'utilisateur. Nous proposons ensuite un ensemble de critères nous permettant d'évaluer, de comparer et de classifier les différentes IUI en fonction des tâches, des utilisateurs…

Notre approche est présentée dans le cadre spécifique des interfaces de visualisation de résultats de recherche. Ces résultats peuvent provenir de moteurs de recherche (cadre applicatif privilégié dans les exemples de la section 2) ou plus généralement de systèmes de recherche d'information. Cependant notre approche reste adaptée à d'autres cadres applicatifs similaires tels que la visualisation exploratoire de données ou la fouille visuelle de données.

2 Exemples d'Interfaces Utilisateur d'Information

De nombreuses IUI ont déjà été proposées. Certaines sont utilisées alors que d'autres n'ont pas dépassées le stade du prototype. Cependant le problème majeur reste les trop grandes disparités entre ces différentes interfaces, ce qui rend toute comparaison impossible. Afin d'illustrer notre propos, nous citons ci-après quelques exemples d'IUI majoritairement proposées dans le cadre de la RI sur le Web. La plus classique et la plus utilisée est celle proposée par GOOGLE[1]. Il s'agit d'une interface (**voir figure 1**) qui propose un affichage linéaire des résultats de recherche sous forme d'une liste triée selon un critère de « pertinence ». VIVÍSIMO[2] reprend ce principe mais en y ajoutant une catégorisation automatique des résultats dans une hiérarchie de répertoires significatifs, *via* une technique de *clustering* à la volée. L'interface obtenue (**voir figure 2**) est alors un ensemble de listes toujours triées selon un critère de « pertinence ». L'interface UJIKO[3] est un autre exemple de catégorisation automatique des résultats (**voir figure 3**). Cet aspect classification des résultats est aussi repris dans des interfaces 2D, telles que GROKKER[4] qui organise les résultats dans des catégories (**voir figure 4**). D'autres interfaces proposent des visualisations sous forme cartographique, généralement basées sur les relations qui existent entre les différents résultats. C'est le cas du méta-moteur de recherche KARTOO[5] qui présente les résultats sous forme cartographique (**voir figure 5**). Dans cette interface l'utilisateur peut visualiser les liens entre les résultats et certains « mots-clés ». Cependant l'aspect linéaire persiste étant donné que les résultats sont présentés sous la forme d'une succession de cartes. D'autres interfaces vont plus loin en exploitant la possible spatialisation du Web. Cela se retrouve notamment dans des IUI basées sur l'affichage des résultats dans un environnement virtuel en 3D, telles que VIOS[6] (**voir figure 6**). Ce concept est aussi utilisé dans le prototype SMARTWEB (Bonnel et al., 2005a,b) qui représente les résultats dans une ville virtuelle en 3D (**voir figure 7**). Mais, contrairement à VIOS, SMARTWEB organise les résultats dans cet espace 3D grâce au calcul d'une carte auto-organisatrice, ce qui permet de regrouper (et de placer) les résultats selon la distribution des mots et ainsi d'avoir une « proximité sémantique ». Il existe également d'autres approches, notamment avec des visualisations 3D.

[1] http://www.google.com
[2] http://www.vivisimo.com
[3] http://www.ujiko.com
[4] http://www.grokker.com
[5] http://www.kartoo.com
[6] *Visual Internet Operating System*, http://computer.howstuffworks.com/vios.htm

Par exemple, CAT-A-CONE (Hearst et Karadi, 1997) utilise une visualisation 3D d'un arbre (*Cone Trees* (Robertson et al., 1991)) afin d'afficher simultanément les résultats obtenus et une hiérarchie de catégories prédéfinies (**voir figure 8**). Un autre exemple est le prototype NIRVE (Cugini et al., 2000) qui permet à l'utilisateur de visualiser et manipuler un ensemble de documents résultant d'une requête sur un moteur de recherche. Il possède plusieurs méthodes de visualisation dont un exemple est proposé sur la **figure 9**. Dans cet exemple, appelé « *Spoke and Wheel Model* », l'utilisateur peut agréger des mots-clés afin de former un plus petit ensemble de concepts. À partir de ces concepts, des groupes de documents sont créés et affichés de telle façon que la distance angulaire entre les groupes soit proportionnelle à la distance logique entre eux. De même les distances radiales entre les documents reflètent les distances métriques qui séparent leurs profils (basés sur les concepts définis par l'utilisateur). On peut également citer d'autres interfaces telles que EASY-DOR (Chevalier, 2002) qui utilisent un affichage 3D basé sur des axes représentant chacun un ou plusieurs mots-clés de la requête (**voir figure 10**).

FIG. 1 – *GOOGLE*.

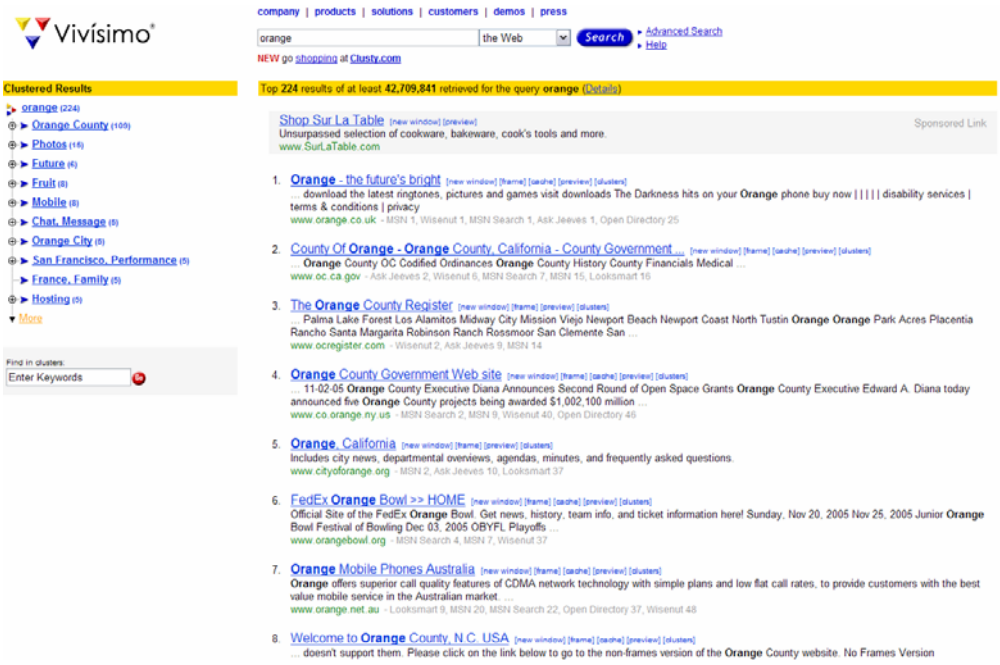


FIG. 2 – *VIVÍSIMO.*

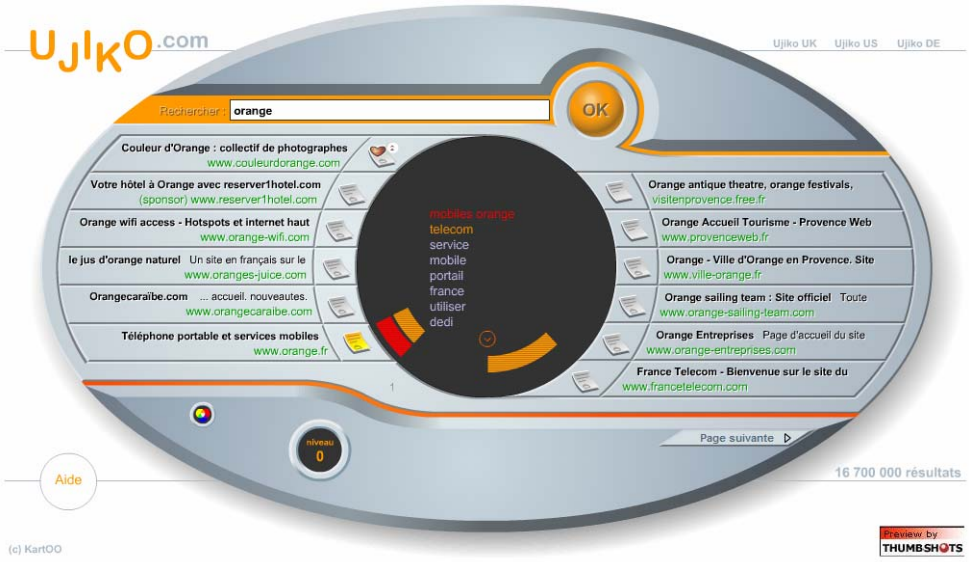


FIG. 3 – *UJIKO.*

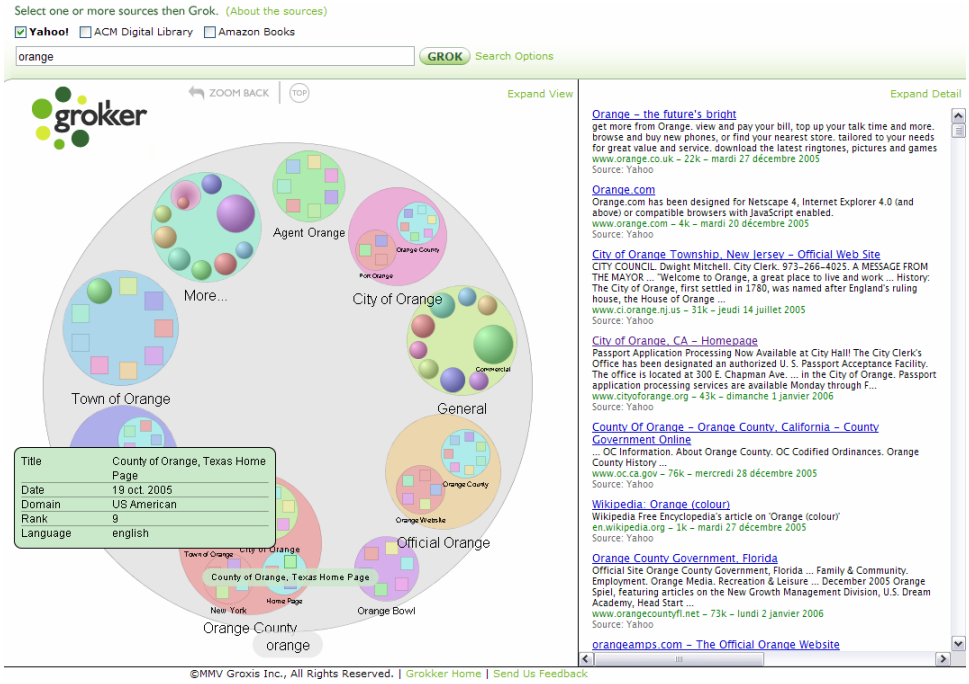


FIG. 4 – *GROKKER.*

FIG. 5 – *KARTOO.*

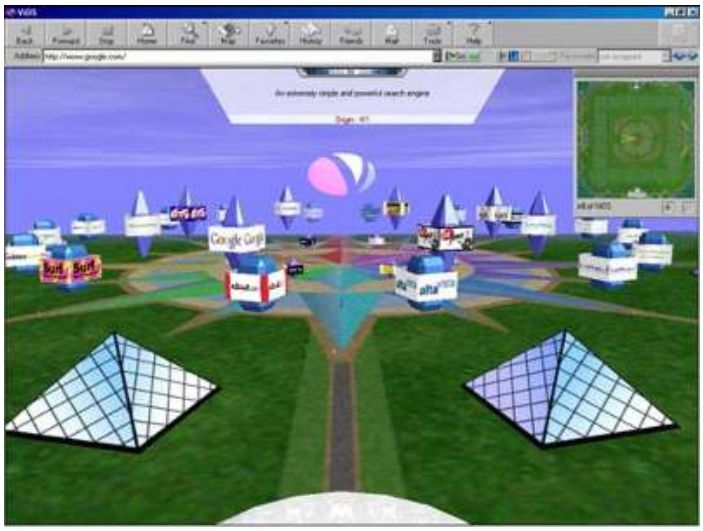


FIG. 6 – *ViOS*.

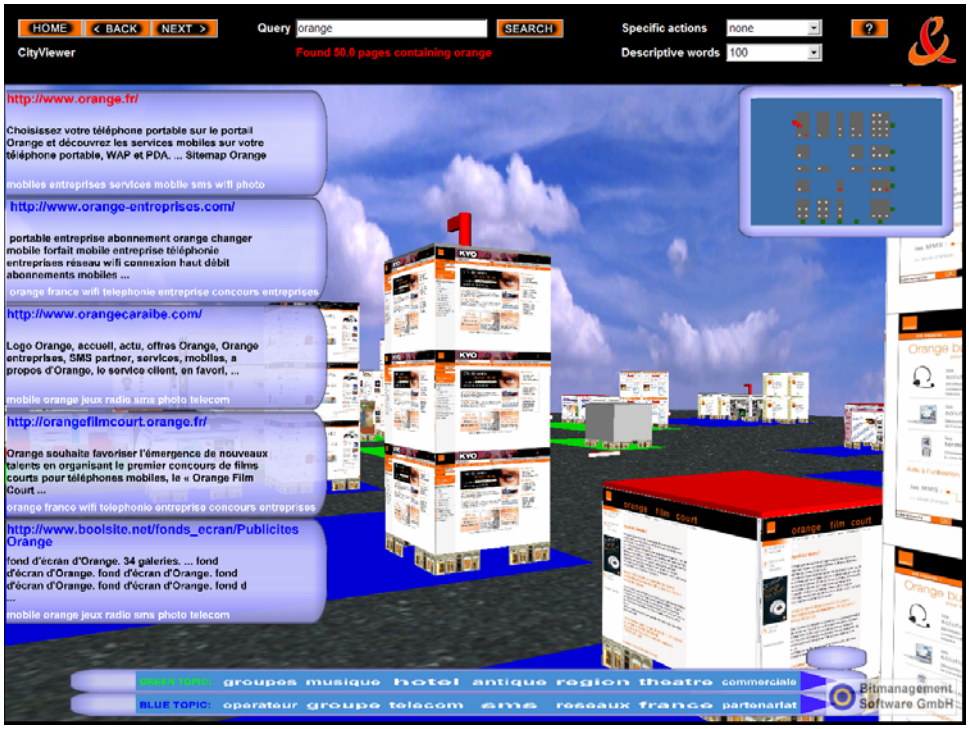


FIG. 7 – *SMARTWEB*.

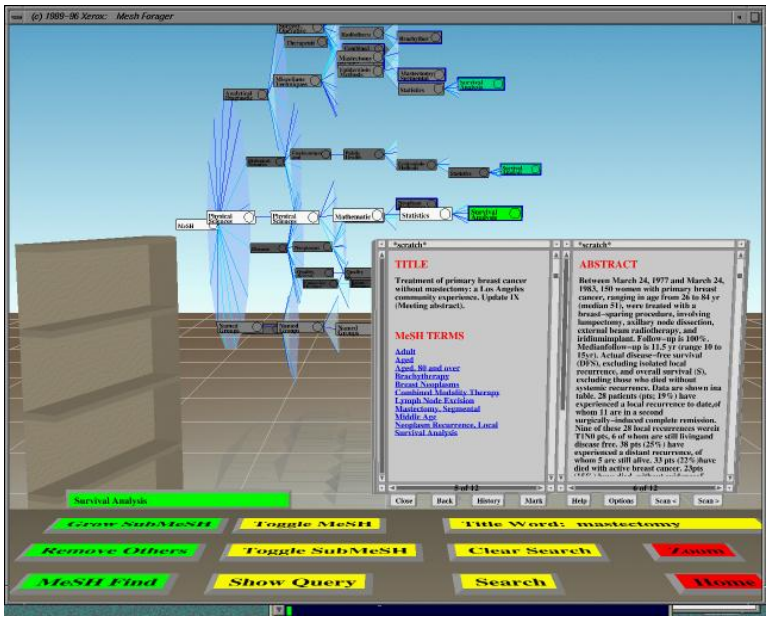


FIG. 8 – *CAT-A-CONE*.

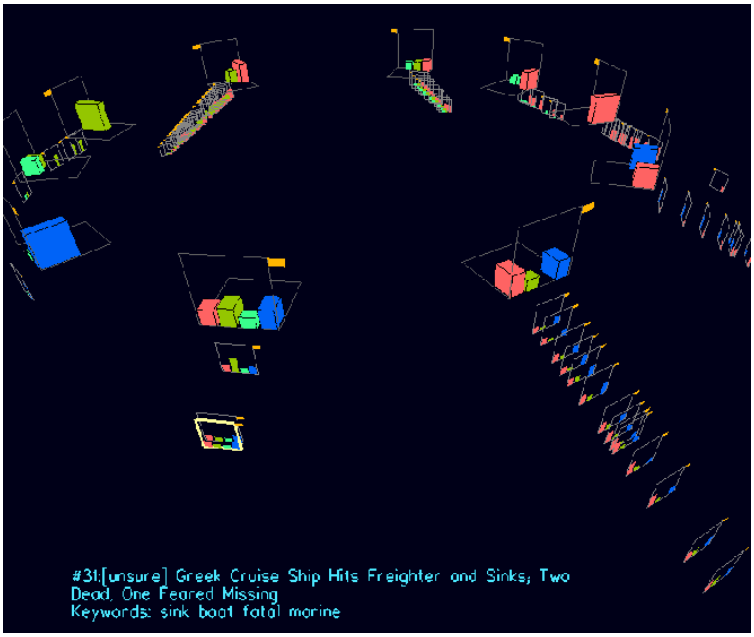


FIG. 9 – *NIRVE, Spoke and Wheel Model*.

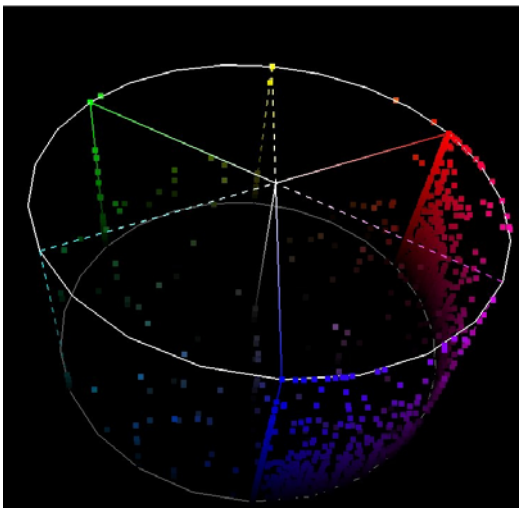


FIG. 10 – *EASY-DOR.*

Toutes ces IUI ne reposent pas sur le même système de recherche et ne proposent pas les mêmes traitements (tels que le *clustering* ou le filtrage des résultats). Aussi, bien que des évaluations isolées de ces IUI soient possibles, elles sont généralement peu interprétables et il est impossible de les comparer de façon efficace, surtout au regard de la tâche de RI. Nous proposons donc dans la suite de cet article de s'intéresser aux conditions nécessaires pour la mise en place d'un cadre expérimental d'évaluation et de comparaison de ces IUI. Nous distinguons alors les contraintes (**section 4**) qui doivent être prises en compte lors de la phase d'évaluation (mais aussi lors de la conception de ces interfaces), et les critères d'évaluation (**section 5**) qui serviront à l'évaluation et à la comparaison avec d'autres systèmes de visualisation. Mais, dans un premier temps, nous nous intéressons au cadre général d'évaluation d'une IUI (**section 3**).

3 Évaluation d'une IUI

La méthode la plus courante pour évaluer une IUI est la réalisation d'une étude utilisateur. Cette technique consiste à proposer un questionnaire aux utilisateurs afin qu'ils répondent à chaque question généralement par une note. Ce questionnaire est très souvent inspiré des propositions de Shneiderman (1998) en matière d'évaluation d'interfaces utilisateur et est très orienté « utilisabilité ». Ce principe a notamment été adopté dans (Chevalier, 2002) et (Bonnel et al., 2005c). Cependant, il s'avère que les résultats de ces tests, dans le cas de l'évaluation d'une IUI, sont difficilement interprétables. En effet, les questions posées ne sont pas toujours adaptées au contexte des IUI et il n'est pas possible de vraiment apprécier l'impact de l'IUI sur les performances (du point de vue de l'usager) du processus de recherche d'information (Chevalier et Hubert, 2005).

Dans le cadre de l'évaluation des IUI, deux approches peuvent être distinguées : l'évaluation d'une IUI utilisant son propre système de recherche (*i.e.* ses propres résultats de

recherche) et l'évaluation d'une IUI indépendante de tout système de recherche. Dans le premier cas, nous pouvons retrouver les principaux *benchmarks* qui ont été mis en place dans le contexte de la recherche d'information, et notamment ceux rattachés aux campagnes d'évaluation TREC[7], CLEF[8] ou encore INEX[9]. Or ces campagnes d'évaluation n'étaient pas initialement prévues pour évaluer l'interface utilisateur, ce qui explique l'absence de certains critères spécifiques aux IUI. De plus, dans ces campagnes d'évaluation, les IUI sont évaluées en même temps que le système de recherche propre à chaque participant. Ainsi, même si l'interaction est mise en évidence, l'impact de l'IUI est noyé dans les performances intrinsèques du système de recherche. Cette façon d'évaluer ne permet pas, en soi, de définir si une IUI est performante ou non pour les tâches auxquelles elle est destinée. Ce type d'évaluation suppose que l'intérêt de l'IUI est corrélé à la performance du système. La question qu'il faudrait alors se poser est : comment se comporterait une bonne interface avec un système de recherche peu performant ? Cette question peut faire sourire et nous pensons tous avoir la réponse. Or, nous pouvons penser que l'interface peut « relever » les performances du système. Par exemple, une IUI basée sur une classification des résultats peut proposer des classes tout à fait pertinentes même si le système ne fait pas de la haute précision. Par ailleurs, une comparaison avec d'autres systèmes de recherche reste « possible » mais elle ne permet pas d'identifier les raisons d'un succès ou d'un échec de l'IUI (est-ce lié au système de recherche ou à l'interface ?). Outre ces campagnes d'évaluation orientées RI, nous pouvons souligner l'existence de campagnes d'évaluation plus orientées interaction homme-machine telles que (Fekete, 2004), qui fournissent une vision complémentaire de l'évaluation d'une IUI « en contexte ». Nous regrettons cependant que ces différentes campagnes d'évaluation n'aient pas plus « d'interactions » entre elles.

Dans cet article, nous nous intéressons donc essentiellement au deuxième cas, c'est-à-dire celui où l'évaluation des IUI peut se faire indépendamment du système de recherche. Si jusqu'à présent les systèmes « les plus chanceux » avaient le privilège d'être évalués, les comparaisons étaient quasi-inexistantes du fait de la trop grande hétérogénéité des systèmes. Dans un premier temps, ce n'est pas tant la comparaison intrinsèque qui nous intéresse mais plus l'évaluation des IUI en prévision de les classifier afin d'identifier, selon le contexte d'utilisation, quelles IUI sont les plus adéquates. Cependant, le cadre d'évaluation de certains critères proposés dans la section 5, peut rendre possible la comparaison de différentes IUI. Par ailleurs, nous sommes convaincus que pour qu'un système de recherche rende les meilleurs services aux usagers, il est nécessaire qu'il propose différentes IUI. Le choix de l'IUI peut alors être, soit proposé explicitement à l'utilisateur qui choisit en fonction de la tâche à réaliser, soit pris en compte dans l'adaptabilité automatique ou semi-automatique du système par rapport aux données à visualiser (ou aussi par rapport à sa connaissance du profil de l'utilisateur). Plusieurs IUI pour un système de recherche permettent alors de répondre aux besoins aussi variés que le sont les usagers. En effet, Shneiderman (1998) soulignait qu'une interface peut ne pas répondre à toutes les tâches et à tous les besoins des usagers. Nous pouvons ainsi faire apparaître que l'utilité d'une IUI dépend donc, en partie, de différents critères liés à l'usager. Pour catégoriser ces critères nous pouvons reprendre la vision de Lainé-Cruzel (1999) qui souligne que l'usager peut être caractérisé par trois questions : Qui

[7] Text REtrieval Conference - http://trec.nist.gov
[8] Cross-Language Evaluation Forum - http://clef.isti.cnr.it
[9] Initiative for the Evaluation of XML Retrieval - http://inex.is.informatik.uni-duisburg.de/2005/

est-il ? Que veut-il ? Pour quoi faire ? Ce ne sont pas les seuls critères qui doivent entrer en compte dans l'évaluation. En effet, il y a certains critères liés aux documents retrouvés en résultat. Ainsi, la visualisation en liste de résultats (*e.g.* GOOGLE) peut suffire pour une tâche donnée et/ou pour un nombre très restreint de documents. Avant de présenter les critères que nous avons identifiés, nous nous intéressons aux différentes contraintes que doit respecter, à notre avis, la phase d'évaluation.

4 La phase d'évaluation

Pour obtenir des résultats exploitables, il est nécessaire de préparer avec soin la phase d'évaluation. Notamment, afin de pouvoir évaluer et comparer efficacement différentes IUI, un certain nombre de contraintes doivent être respectées lors de la phase d'évaluation. Ces contraintes doivent être prises en compte dès la conception de l'interface afin de pouvoir garantir leur respect. Ainsi l'IUI doit être :

- **indépendante** du système de recherche (*i.e.* des données). Il est en effet nécessaire que l'IUI accède aux données via des interfaces proposées par les systèmes de recherche et libres d'utilisation ;
- **modulable** en terme de fonctionnalités, c'est-à-dire que les fonctionnalités doivent être adaptables au système de recherche utilisé. En effet, l'interface peut être destinée à plusieurs outils ou applications ;
- **capable** de gérer des jeux de requêtes prédéfinis.

Ces différents aspects sont présentés dans les sous-sections suivantes.

4.1 Données

Le premier point à vérifier est de s'assurer que les interfaces traitent le même ensemble de résultats, c'est-à-dire que les données visualisées sont identiques. Dans notre cas, les résultats peuvent être déclarés identiques à condition de respecter trois égalités : celle du contenu, celle du nombre et celle des descripteurs disponibles (*i.e.* les mêmes informations de base sur les documents sont disponibles mais chaque IUI peut les traiter différemment). La solution la plus simple pour répondre à cette contrainte fondamentale à l'évaluation et à la comparaison de deux IUI, est d'utiliser les API[10] fournies par certains moteurs de recherche. Un système de recherche de référence pourrait être proposé si un moteur de recherche proposait une API suffisamment riche en termes de fonctionnalités. Malheureusement, les API existantes ne donnent pas accès à l'ensemble des descripteurs disponibles dans les bases d'index. Une solution pourrait alors être la proposition d'un système de recherche destiné à l'évaluation des IUI et fournissant une API riche. Il faudrait que l'API propose des fonctions d'accès suffisamment riches pour que chaque IUI y trouve son compte. Cependant, des besoins supplémentaires de la part d'une IUI peuvent exister, ce qui se traduirait par une différence en ressources nécessaires. Dans ce cas, il faudrait aussi que l'API soit suffisamment ouverte pour la faire évoluer et y intégrer ces nouveaux besoins afin d'améliorer le cadre d'évaluation et de comparaison de ces interfaces.

[10] Application Programming Interface

4.2 Fonctionnalités

Il est nécessaire que les différentes IUI évaluées puissent s'adapter en fonction de l'API proposée ; les fonctionnalités de ces IUI étant très souvent dépendantes des informations fournies par l'API. Ainsi, en supposant que les moteurs de recherche adoptent une API partagée, l'utilisation d'une IUI avec n'importe quel moteur sera désormais possible. Ce n'est pas le cas aujourd'hui est cette vision reste à l'heure actuelle utopique.

4.3 Jeux de requêtes

Il est important que chaque IUI exploite les mêmes jeux de requêtes permettant de couvrir les différentes tâches que l'on souhaite évaluer. Par exemple, en recherche d'information, nous pourrions identifier les tâches suivantes :

- recherche d'un élément précis (exemple : existe-t-il une page web professionnelle de François Poulet ?). En général le résultat peut être assimilé à une information booléenne dans le sens où l'usager cherche une information très précise ;
- recherche de l'existence d'une information *a priori* connue (exemple : je recherche les sites web parlant des lois de Murphy) ;
- recherche exploratoire, c'est-à-dire panorama d'un thème donné (exemple : je recherche tout ce qu'il existe dans le domaine de la gestion des connaissances).

Chacune de ces tâches n'implique pas, par leur nature, la même restitution en termes d'IUI. Les tâches doivent donc être définies avant la phase d'évaluation car les jeux de requêtes devront répondre aux différentes tâches identifiées que l'on souhaite évaluer. De plus, chacune des tâches identifiées devra être caractérisée par un ou plusieurs critères (définis en section 5) qui pourront être éventuellement pondérés. Par exemple, nous pouvons imaginer que la tâche « recherche d'un document précis » soit caractérisée entre autres par un temps de réussite de la tâche très court. Ainsi le poids de ce critère pourra être prédominant par rapport aux autres (*i.e.* critère pondéré avec un poids fort).

4.4 Les évaluateurs

Le choix des « évaluateurs » devra également être réalisé avec soin pour obtenir un panel d'usagers représentatif des catégories d'usagers pour lesquelles les IUI vont être testées. Ainsi, les usagers pourront par exemple être répartis en neuf classes selon les caractéristiques des usagers en RI (*c.f.* tableau 1). Ce classement peut être réalisé par l'intermédiaire d'un questionnaire ouvert par exemple.

		Connaissance du domaine de recherche		
		Néophyte	Intermédiaire	Expert
Connaissance de l'outil informatique (2D/3D…)	Néophyte			
	Intermédiaire			
	Expert			

TAB. 1 – *Catégories des usagers des IUI (échelle ouverte).*

Les évaluations pourront alors se dérouler avec des jeux de test spécifiques (à la tâche) et identiques pour tous les évaluateurs. À la suite des différentes évaluations, les résultats pourront être synthétisés pour vérifier l'adéquation entre l'IUI, les tâches et les usagers. Ceci peut être réalisé par le calcul de corrélations par exemple.

5 Critères d'évaluation

Dans cette section, nous proposons quelques critères d'évaluation permettant de comparer des IUI respectant les contraintes précédemment citées. La liste de critères proposée n'est pas exhaustive et certains critères peuvent ne pas être adaptés à toutes les interfaces. L'idée étant de faire émerger des relations entre des critères et la réussite ou non d'une recherche d'information. La sélection des critères à prendre en compte doit être effectuée en amont de la phase d'évaluation et doit correspondre à une caractérisation des différentes tâches évaluées. En effet, l'évaluation d'une IUI, pour une tâche donnée, se traduit par la prise en compte d'un certain nombre de critères et éventuellement par une pondération de ces critères.

5.1 Temps nécessaire pour achever la tâche

Ce critère consiste à mesurer le temps mis par l'utilisateur pour achever la tâche qu'il doit réaliser. Ainsi ce temps peut être par exemple le temps nécessaire pour trouver un résultat pertinent pour la requête effectuée. Ce critère est donc mesuré objectivement mais nécessite d'avoir préalablement identifié l'ensemble des résultats pertinents pour la requête. Il peut également s'agir de mesurer, pour certaines tâches, le temps nécessaire à l'usager pour identifier un ensemble (ou la totalité) de documents pertinents (similaires ou non). Dans ce cas, le nombre de résultats à retrouver par l'utilisateur doit être fixé et cohérent par rapport au nombre total de résultats pertinents pouvant être retrouvés. On peut également procéder inversement, en fixant le temps de recherche et en comptant le nombre de résultats pertinents identifiés par l'utilisateur. La mesure de ce critère reste alors toujours objectivement quantifiable.

5.2 Ressources nécessaires pour remplir une tâche

Le temps pour trouver le résultat n'est pas forcément pertinent surtout si l'on considère des usagers peu habitués aux applications informatiques. Ainsi, le nombre d'actions à réaliser ainsi qu'un malus dans le cas d'un retour en arrière pourrait être couplé au temps mis pour achever la tâche.

5.3 Temps pour mémoriser et retourner à un document précédemment identifié

Ce critère mesure le temps nécessaire à l'utilisateur pour qu'il retrouve un document qu'il a précédemment identifié (ou qu'il connaît) dans l'interface proposée. Il peut s'agir d'un résultat que l'on vient de voir dans l'espace des réponses d'une requête (*i.e.* on n'a pas quitté

l'espace des réponses) ou d'un document dont on sait qu'il appartient aux résultats d'une requête (*i.e.* on a quitté l'espace des réponses pour y revenir plus tard).

5.4 Temps de réponse du système

Il s'agit là d'un critère simple mais inévitable. En effet, la rapidité est une notion fondamentale pour l'utilisateur lors d'une tâche de recherche d'information. Le calcul du temps de réponse du système consiste simplement à mesurer le temps écoulé entre la validation de la requête et l'affichage des résultats. On considère ici que l'affichage des résultats correspond à l'affichage du dernier composant graphique de l'interface. Cependant, il n'est plus suffisant de se contenter d'une seule valeur. En effet, les traitements des IUI sont de plus en plus variés. Les deux traitements les plus fréquents sont généralement l'affichage et le calcul d'un *clustering* à la volée. Dans ce cas, il semble important d'être en mesure de différencier le temps de calcul pour ces deux tâches. Cette différence permet d'effectuer une meilleure comparaison des systèmes et d'identifier les raisons de leurs potentielles lacunes.

5.5 Compréhension ou « utilisabilité » de l'interface

Il s'agit d'un critère subjectif pouvant être évalué par des questions posées à l'utilisateur sur le fonctionnement de l'interface (Shneiderman, 1998). Cependant, afin de garantir « l'égalité » des évaluations par rapport à ce critère, les mêmes conditions doivent être respectées : temps d'explication de l'interface, temps de manipulation de l'interface, présence d'une aide en-ligne. Dans le cas des IUI, le succès d'une interface passe bien souvent par une compréhension intuitive de l'interface (*i.e.* sans explications pour les fonctionnalités fondamentales). On peut aussi ne pas chercher à évaluer ce critère, considérant qu'il influe plus ou moins dans l'évaluation d'autres critères tels que le temps de recherche de résultats pertinents.

5.6 Capacités de personnalisation et d'adaptation

Dans un premier temps, ce critère peut-être évalué de manière binaire : l'interface est-elle personnalisable et/ou est-elle adaptative ? La personnalisation concerne la possibilité offerte à l'utilisateur de modifier certains paramètres de l'interface alors que l'aspect adaptatif est relatif à la prise en compte par l'IUI de certaines connaissances qu'elle possède de l'usager. Dans le cas où l'IUI est personnalisable, les points suivants pourraient alors être considérés : quelles parties (ou fonctions) sont personnalisables, quelles compétences sont nécessaires du point de vue de l'utilisateur, quel est le temps nécessaire pour l'utilisateur pour réaliser ces personnalisations, quel est l'apport de la personnalisation... Concernant le critère d'adaptation, on pourrait entre autres vérifier si l'interface exploite les précédentes requêtes formulées par l'utilisateur pour adapter, réorganiser les résultats comme le propose par exemple Lainé-Cruzel (1999). Un autre point consiste à vérifier si l'IUI est en mesure de présenter la représentation de l'usager qu'elle possède. Ce dernier point est prédominant pour une bonne acceptation de la personnalisation proposée. Et là aussi il faudrait pouvoir identifier l'apport de l'adaptation (le résultat est-il plus pertinent ?...). Ce critère de personnalisation et d'adaptation de l'IUI reste cependant trop dépendant de nombreux aspects

(tels que du type d'interface ou de personnalisation proposés) pour tenter une généralisation de l'évaluation de ce critère.

5.7 Réussite de la tâche proposée

L'utilisateur doit juger s'il a réussi ou non à réaliser la tâche qui lui était proposée. Cet achèvement peut être binaire ou être associé à une échelle ouverte (note de 1 à 5 par exemple). L'intérêt de ce critère est qu'il permettra de calculer l'adéquation entre l'IUI, la tâche et le type d'utilisateur par exemple.

5.8 Caractéristiques du résultat de recherche

Les critères à prendre en compte ne sont pas tous uniquement liés à l'usager. En effet, pour une tâche donnée, une IUI peut être efficace pour un nombre limité de documents retrouvés et ne plus l'être pour un nombre plus important. Ainsi, nous pouvons souligner les caractéristiques suivantes :

- nombre de documents retrouvés ;
- longueur des documents (en moyenne et en écart type) ;
- hétérogénéité des thèmes abordés dans les documents ;
- hétérogénéité du contenu des documents (contiennent-ils du texte, des images…) ;
- hétérogénéité de structure (est-ce des documents XML, du texte libre ou des données structurées ?).

Ces différents critères devront être combinés pour mesurer l'adéquation entre l'IUI, la tâche de l'usager et le type d'individu. En effet, les usagers peuvent être caractérisés par deux caractéristiques spécifiques pour la recherche d'information : la connaissance du domaine et la connaissance pratique (des outils informatiques). Une étude souligne l'importance de ces critères qui influent sur la façon dont les utilisateurs recherchent l'information (Hölscher, 2000). Grâce à ces différentes informations nous pouvons donc s'assurer qu'une IUI s'adapte à certains publics et à certaines tâches.

6 Conclusion

Il existe un grand nombre d'outils de recherche d'information notamment disponibles sur le Web. Ces différents outils tentent de répondre au mieux aux attentes des usagers aussi variés que les besoins en information qu'ils tentent d'assouvir. La stratégie actuelle est de faire progresser les outils en termes de fonctionnalités intrinsèques (amélioration du processus d'indexation, adaptation…), mais un aspect important, lié à l'usager, reste la conception et l'intégration de l'IUI. Cette IUI est la passerelle directe entre l'outil de recherche et l'usager au travers de laquelle les documents retrouvés sont présentés. Ces IUI doivent ainsi être bien réfléchies pour permettre à une catégorie d'usagers de remplir des tâches spécifiques. Cependant, à l'heure actuelle nous ne sommes pas en mesure de vérifier si au moins une IUI répond bien aux attentes. Le besoin en évaluation est donc important pour

vérifier ces hypothèses mais également de vérifier si éventuellement une IUI pouvait remplir d'autres tâches non prévues initialement ou si elle était adaptée à d'autres catégories d'usagers. Par ailleurs, cette évaluation permet en outre de réaliser une comparaison et une classification des IUI en fonction des tâches qu'elles permettent de remplir par exemple. Nous proposons dans ce contexte une approche de l'évaluation des IUI au regard de différents critères. Ces critères permettent la caractérisation des tâches que l'on souhaite évaluer. Grâce aux évaluations et aux mesures ainsi calculées au travers des différentes évaluations, une synthèse pour chaque IUI pourra être fournie. Cependant, la phase d'évaluation n'est pas si simple à réaliser et surtout à organiser comme le souligne par exemple Fekete (2004). Nous devons désormais poursuivre notre réflexion sur les indicateurs et la caractérisation des différentes tâches à partir de ceux-ci. Nous devons également approfondir le mode opératoire de cette phase d'évaluation afin d'obtenir et de valider les résultats que nous espérons aussi objectifs que possible.

Références

Bonnel N., Cotarmanac'h A., et Morin A. (2005a). Gestion et visualisation des résultats d'une requête. *Actes du 3ème atelier Visualisation et Extraction de Connaissances (associé à EGC'05)*, pp. 37–47.

Bonnel N., Cotarmanac'h A., et Morin A. (2005b). Meaning Metaphor for Visualizing Search Results. *Proceedings of the 9th International Conference on Information Visualisation*, pp. 467–472, IEEE Computer Society.

Bonnel N., Cotarmanac'h A., et Morin A. (2005c). Visualisation 3D des résultats de recherche : quel avenir ? *Créer, jouer, échanger : expériences de réseaux (Actes de la conférence H2PTM'05 : Hypermedias Hypertexts, Products, Tools and Methods)*, pp. 325–339, Hermes Science Publications.

Chevalier M. (2002). *Interface adapative pour l'aide à la recherche d'information sur le web*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France.

Chevalier M. et Hubert G. (2005). Evaluation d'une interface de restitution de recherche : Quelles conclusions en tirer ? *Actes du 3ème atelier Visualisation et Extraction de Connaissances (associé à EGC'05)*, pp. 15–27.

Cugini J., Laskowski S., et Sebrechts M. (2000). Design of 3D Visualization of Search Results: Evolution and Evaluation. *Proceedings of IST/SPIE's International Symposium: Electronic Imaging 2000: Visual Data Exploration and Analysis*.

Fekete J.-D. et Plaisant C. (2004). Les leçons tirées de deux compétitions de visualisation d'information. *ACM IHM 2004*, pp. 7–12.

Hearst M. et Karadi C. (1997). Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy. *Proceedings of International ACM/SIGIR Conference*, pp. 246–255.

Hölscher C. et Strube G. (2000). Web search behavior of internet experts and newbies. *Proceedings of the 9th International Conference on the World Wide Web (www9)*.

Lainé-Cruzel S. (1999). ProfilDoc – Filtrer une information exploitable. *Bulletin des bibliothèques de France (BBF)*, 44(5), pp. 60–64.

Robertson G., Mackinlay J., et Card S. (1991). Cone Trees: Animated 3D Visualizations of Hierarchical Information. *Proceedings of ACM CHI'91 Human Factors in Computing Systems Conference*, pp. 189–194, ACM Press New York.

Shneiderman B. (1998). *Designing the User Interface*. Addison-Wesley.

Annexe : récapitulatif

Cette annexe permet de faire un récapitulatif des contraintes et critères proposés. Le tableau 2 constitue en quelque sorte un mémo d'évaluation, récapitulant un ensemble de contraintes à satisfaire afin de pouvoir utiliser certains des critères proposés en vue d'une comparaison avec d'autres systèmes.

CONTRAINTES
1. Données
2. Fonctionnalités
3. Jeu de requêtes

PRÉALABLES
1. Définir et caractériser les tâches à évaluer
2. Choisir un panel d'usagers représentatif

CRITÈRES
1. Temps nécessaire pour achever la tâche
2. Ressources nécessaires pour accomplir une tâche
3. Temps pour mémoriser et retourner à un document précédemment identifié
4. Temps de réponse du système
5. Compréhension ou utilisabilité de l'interface
6. Capacités de personnalisation et d'adaptation
7. Réussite de la tâche proposée
8. Caractéristiques du résultat de recherche

TAB. 2 – *Récapitulatif.*

Summary

An information retrieval (IR) process has no sense for users without the last step which consists in visualizing the search results. The increase of the importance of search result visualization is responsible of the proposition of many interfaces in the last few years. These interfaces can be based on textual, 2D or 3D metaphors. Although some of these interfaces were evaluated, no effective comparison was carried out in this context due to the lack of constraints on these interfaces and due to the lack of comparison criteria concerning the IR task. So we propose to introduce some evaluation tracks in order to obtain a framework enabling the evaluation and the comparison of various **Information User Interfaces**.

Keywords: Information User Interface, Visualization of Search Results, Information Retrieval System, Evaluation and Comparison Criteria.

Visualisation de navigation sur interface graphique pour l'analyse cognitive de parcours

Marc Damez*, Stephan Renaud**

* Université Pierre et Marie Curie-Paris6, UMR 7606, DAPA, LIP6, 8 rue du Capitaine Scott
F-75015 Paris, France
SEJER, 30 place d'Italie 75702 Paris Cedex 13
Marc.Damez@lip6.fr

** Laboratoire Cognition & Usages,
Université de Paris 8, 2 rue de la liberté, 93626, St Denis, France
Stephan.Renaud@cognition-usages.org

Résumé. Les modélisations utilisateurs telles qu'elles sont étudiées par les informaticiens ou les psychologues cognitivistes manipulent souvent des données très abondantes et possèdent peu d'outils de fouille visuelle. Une visualisation des navigations sur une interface graphique pour l'analyse cognitive de parcours est présentée ici et deux exemples d'interprétation y sont proposés.

1 Introduction

Les analyses des navigations d'utilisateurs sur une interface multimédia sont souvent utilisées pour la conception des interfaces adaptatives, la classification de comportements, la détection d'actions critiques, les systèmes d'aides automatiques, l'évaluation ergonomique d'interfaces, etc. Ces analyses utilisent généralement différentes techniques de statistiques, d'apprentissages artificiels ou d'analyses exploratoires de données permettant la mise en évidence des aspects cognitifs des actions utilisateurs. Cet article se place dans cette dernière catégorie et présente une méthode de visualisation de données provenant des évènements générés par une interface graphique.

Naturellement présents dans tout type d'interface, ces évènements sont très utiles dans la conception d'automates car ils constituent une véritable trace de l'interaction homme-machine. Les représentations graphiques de ces données sont difficiles à mettre en œuvre, tant à cause de la profusion d'évènements générés, que par les types d'analyses à réaliser. Aussi, en s'inspirant des spécifications UML scénario (Rumbaugh et al., 2004), Damez (2006) a proposé une visualisation de ces données permettant une analyse cognitive d'une interactions homme-machine.

Les analyses cognitives de parcours peuvent porter sur de nombreux aspects de l'utilisation d'une interface. En particulier, les Nouvelles Technologies d'Information et de Communication (NTIC) étudient généralement les effets et les implications des nouveaux dispositifs et logiciels. Pour ces analyses, les psychologues utilisent toutes les sources d'informations possibles (plans de caméra, audio, logiciel de traces). Les traces permettent les analyses les plus fines, mais elles sont encore trop peu exploitées. Une des contraintes des

psychologues est de manipuler ces grandes quantités d'informations pour repérer des phénomènes isolés ou répétés et de s'en faire une représentation. Ils ont besoin d'outils d'analyse de parcours souples et précis qui offrent une représentation visuelle.

Pour répondre a de tel besoins, nous présentons ici un logiciel de visualisation que nous avons mis au point et deux exemples d'analyses : l'analyse de parcours sur différents types d'interfaces hypermédia d'un manuel scolaire électronique et l'analyse de parcours d'un groupe d'utilisateurs pour mettre en évidence des comportements en fonction des contextes d'utilisation.

2 La visualisation de navigation sur interface graphique

Dans cette partie, nous présentons les besoins des psychologues cognitivistes en terme de besoin de visualisation de données ainsi que l'outil graphique de visualisation de navigation que nous avons conçu.

2.1 Les nécessités pour une analyse cognitive de parcours

Des modèles d'analyses cognitives de parcours tels que CPM-GOMS ont été automatisés dans un cadre d'expérimentation très restreint (John et al., 2002). De la même façon, une automatisation de construction de graphes pour analyser les comportements de navigation sur interface graphique est proposée ici.

Les problématiques des psychologues cognitivistes sont nombreuses car leurs expérimentations génèrent souvent un grand nombre de données à analyser, et les processus automatiques pour les étudier sont souvent nombreux et complexes. Aussi, ont-ils besoin d'outils simples, souples, et facilement interprétables pour mettre en évidence des caractéristiques de navigation. Ainsi, l'analyse d'actions élémentaires, comme une prise de décision simple (un clic de souris), nécessite la remise en contexte par la visualisation du parcours dans son ensemble. De plus, la comparaison des parcours d'un groupe d'utilisateurs permettra l'évaluation des composants de l'interface et l'évaluation de son utilisabilité. La méthode de visualisation proposée possède également l'avantage de permettre de représenter plusieurs types d'interfaces sur un même graphe, et ainsi de comparer les différentes propriétés ergonomiques de chacune.

2.2 L'existant

Des outils tels que « Webviz » (Pitkow et al., 1994) permettent de mettre en évidence des comportements d'internaute. La représentation utilisée nécessite la connaissance a priori de l'information recherchée et, par l'application d'un filtre sur la visualisation, la validation d'une hypothèse peut être réalisée. Par contre, les aspects temporels de la navigation et du comportement des utilisateurs sont difficilement visualisables dans leur ensemble avec cet outil.

Un logiciel commercial comme « Clicktracks » (voir http://www.clicktracks.com) utilise également les évènements de l'interface et représente la probabilité de click sur un lien d'une page Internet. Dans ce cas, il y a une remise en contexte de l'utilisation de l'interface mais sur un instant très précis de la navigation ce qui permet l'étude et la comparaison de l'utilisation de différentes interfaces.

« Wum » (Spiliopoulou et Faulstich, 1998) utilise une représentation arborescente, qualifié d'agrégée car manipulant des statistiques de navigation, pour mettre en évidence des comportements d'utilisateurs. Par contre, « Wum » ne donne que très peu de renseignement sur le contenu de l'interface utilisée.

Une autre approche de modélisation utilisateur pour la recherche d'information sur le web utilisant un outil de visualisation est présentée par Delort (2003). Dans ce cas, l'étude est centrée sur le contenu des pages Internet visitées et des indices de changements de stratégies de navigation par l'utilisateur sont proposés. Cependant la visualisation ne renseigne pas sur le véritable comportement de l'utilisateur, la visualisation n'étant qu'une représentation du parcours des contenus visités.

Notre approche permet la remise en contexte de l'ensemble du parcours sur une interface graphique. Cette visualisation permet également de représenter les éléments de l'interface qui sont manipulés par différents utilisateurs et, par extension, de comparer les différentes utilisations de plusieurs interfaces. Comme cette visualisation est directement construite à partir des traces générées par les évènements de l'interface, les aspects temporels et/ou séquentiels des parcours sont conservés de façon stricte. Une option de notre outil permet également la visualisation de la séquence des tâches courantes de l'utilisateur et donc la remise en contexte de l'utilisation de l'interface.

2.3 L'outil

Les évènements utilisateurs manipulés par le système graphique sont relevés directement dans une structure arborescente décrivant l'interface sur laquelle les actions sont portées. En conservant la notion de parcours temporel, et en s'inspirant des spécifications UML en mode scénario, les actions des utilisateurs sont représentées sur les objets cibles (le temps étant ici représenté sur l'axe horizontal). La figure 1 présente un extrait de visualisation de parcours de plusieurs utilisateurs.

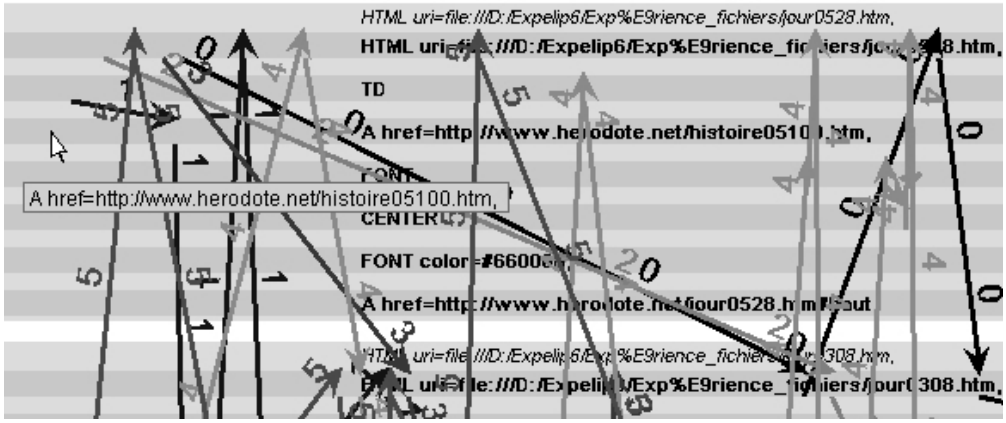


Figure 1. Extrait d'une visualisation

Sur cette figure, les barres horizontales représentent les « Graphical User Interface » (GUI) qui ont fait l'objet de manipulation de la part des utilisateurs. Les barres en gris clair sont les conteneurs (ici les pages Internet « HTML uri=file:///D:/Expelip6/Exp%E9rience

_fichiers/jour0528.htm, » et « HTML uri=file:///D:/Expelip6/Exp%E9rience_fichiers/jour0308.htm, »). Les barres grises foncées sont les cibles (« HTML uri=file:///D:/Expelip6/Exp%E9rience_fichiers/jour0528.htm, », « TD », « A href=http://www.herodote.net/histoire05100.htm, », « FONT », « CENTER », etc.) qui ont fait l'objet d'une manipulation par les utilisateurs. Les flèches numérotées représentent les parcours sur ces cibles à raison d'un utilisateur par couleur de flèche et/ou par numéro. La longueur d'une flèche projetée sur l'axe horizontal représente la durée séparant les débuts des interactions entre l'objet à l'origine de la flèche et l'objet à la fin de la flèche.

Une analyse porte sur une tâche réalisée par un ensemble d'utilisateur. Chaque utilisateur accomplit une séquence de sous-tâches qui peut se décomposer en plusieurs sous-séquences. Un outil de sélection de sous-séquences permet de faire apparaître celles-ci sur le schéma. On peut ainsi comparer les différentes façons qu'ont les utilisateurs pour réaliser une même sous-tâche. Nous avons également incorporé la possibilité de sélectionner l'ordre des GUI apparaissant dans la visualisation. Ceci permet la construction de graphe bien plus lisible (l'exemple de la partie 1 ordonne les GUI de deux interfaces différentes sur le même schéma). Pour les graphes présentés dans les parties suivantes, nous avons choisi de montrer la totalité du parcours, car ils sont plus compréhensibles pour l'interprétation cognitive. Cependant lors de l'utilisation de cette visualisation par des experts cognitivistes, une fonction de zoom pour les traces temporellement longues a été implémentée. Il y a également un outil de sélection de traces visibles, pour l'analyse d'un grand nombre de traces et un outil de normalisation du temps permet de comparer plus facilement le séquençage des actions de l'utilisateur.

3 Eléments d'analyses de parcours sur différents types d'interfaces hypermédia d'un manuel scolaire numérique

3.1 Le contexte et l'outil

Dans le cadre d'un projet d'analyse sémantique des contenus pour un hypermédia pédagogique, un nouveau type de navigation a été conçu. Les navigations dans les manuels scolaires numériques sont encore fortement influencées par la présentation des manuels papiers, elles s'opèrent de page en page selon un parcours organisé par les concepteurs (auteur et éditeur) dans le but de fournir des supports de cours en classe pour le professeur. Ainsi, l'hypertextualité est réduite à l'arborescence des manuels papiers, les documents sont organisés en chapitres et sous-chapitres.

Dans le nouvel outil de navigation, le parcours se fait de document en document, l'organisation de ces derniers étant calquée sur celles des concepts. Cette organisation est obtenue à partir des critères de la typologie des relations conceptuelles décrite par Vignaux [Vignaux 2002]. Trois types de relations ontologiques adaptées pour notre expérience (temporelle, catégorielle et granulaire) organisent les concepts sur 3 axes bijectifs, ce qui permet 6 types de déplacements qui sont à la fois des articulations logiques et langagières. Ainsi, les déplacements dans l'hypermédia sont à la fois des changements de documents et des « pas » conceptuels, l'espace de navigation peut être considéré comme un « espace problème » où chaque déplacement est révélateur d'un type d'activité cognitive de l'utilisateur (compréhension, désorientation, confrontation d'informations etc.). Pour que les déplacements puissent être interprétés, il est nécessaire qu'ils soient contextualisés dans un

parcours, ce qui nous ramène à la visualisation du protocole dans son ensemble pour en saisir la dynamique.

3.2 Expérimentation

Les concepts de 2 chapitres de géologie du manuel de « Sciences de la Vie et de la Terre » de 4ième des éditions Bordas ont été organisé en fonction des critères définis précédemment. Cinq conditions (5 types d'interfaces) ont été testées et une classe d'élèves de 5ème a participé aux tests (les élèves n'ayant ainsi aucune connaissance a priori du concept à apprendre). Les documents pouvaient être des images (C2 et C4) ou des textes (C3 et C5) et les liens propres à notre navigation étaient représentés par les liens textuels (C2 et C3) ou fléchés (C4 et C5). La première condition C1 étant la condition témoin (manuel avec navigation uniquement arborescente).

Document Type de lien	Image	Texte
Texte	C2	C3
Flèche	C4	C5

Tableau 1 Récapitulatif des différentes conditions/expériences effectuées.

Nous avons demandé aux élèves d'expliquer la formation de l'argile. Ils disposaient pour cela de 20 minutes et d'une des 5 versions (C1-5) d'un manuel électronique.

3.3 Résultats

La visualisation des traces de 4 utilisateurs ayant réalisés cette expérience est présentée dans la Figure 2. Sur ce schéma, les barres du haut (en A) représentent les pages visitées par les élèves manipulant la condition 1 (C1). Les barres du milieu (en B), les plus larges, représentent les GUI communs à toutes les versions des manuels. Ce sont les menus, outils, et l'arbre de la structure générale d'un manuel. Les barres du bas de la figure (en C) représentent les pages visitées par les élèves manipulant la condition 3 (C3).

La visualisation de l'ensemble des parcours permet de repérer rapidement (sans traitement, ni analyse des données) les évènements clés, les épisodes et les types de navigation. On voit que la condition C3 est plus efficace que la condition C1 (temps total de réalisation de 4 minutes contre 6) et qu'elle réduit les passages dans l'arborescence (en D). Les participants de la condition C1 terminent le test dans le même temps (en E), en finissant sur la même page qui donne une information non pertinente pour la résolution de l'exercice, ces parcours mènent à l'échec en un temps supérieur à celles des participants des conditions C3. Ces derniers réussissent ensemble en F après une utilisation du nouveau module de navigation découvert en G. A partir de ce moment les traces sont moins accidentées, ce qui montre une utilisation du nouvel outil de navigation. Un participant possède un parcours particulièrement « ondulé » en G, ce qui indique une bonne pratique de la nouvelle navigation acquise en moins de 2 minutes. Le temps très faible de la familiarisation montre la qualité ergonomique de ce nouvel outil. De plus, les temps de lecture se rallongent, ce qui est un indice de la pertinence des pages visitées.

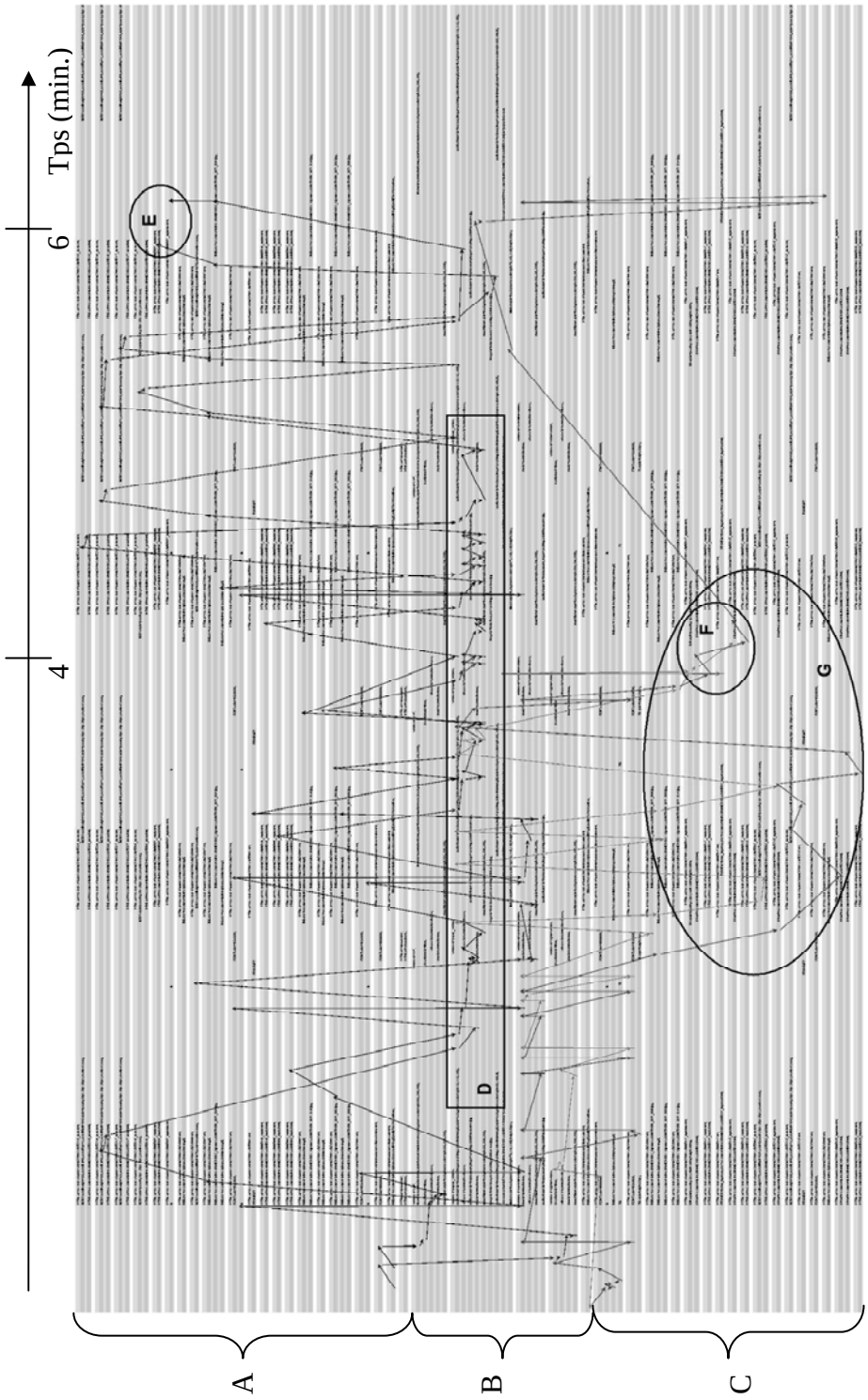


Figure 2. Visualisation des traces de 2 utilisateurs sur 2 interfaces différentes

La visualisation des parcours met en évidence les phénomènes et les résultats remarquables des passations ce qui facilite et accélère l'analyse qualitative. En plus de donner des informations sur la temporalité, les pages visitées et revisitées, elle permet d'étudier les phases des parcours, les types de parcours, les et leurs similitudes (ici les duos des participants présents en E et en F), ce qui permet d'envisager une nouvelle approche des styles cognitifs de navigation. Autre grand intérêt de cette représentation, elle permet de sélectionner les épisodes intéressants pour l'observation des vidéos et les analyses qualitatives les plus fines. Pour les psychologues cognitivistes, l'utilisation de ce logiciel offre un confort d'analyse sans précédent en permettant de multiplier les analyses rapidement.

4 Analyse de comportement et remise en contexte de l'utilisation de l'interface

Dans le cadre de tâches à réaliser sur un hypermédia clos, c'est-à-dire où chaque utilisateur doit impérativement accomplir une action, représentant la fin d'une sous-tâche et le début d'une autre, il est possible de visualiser des différences de comportements dus notamment au contexte créé par la réalisation des sous-tâches précédentes. (Brézillon et al., 2005) ont proposé la représentation de « graphe contextuel » pour la mise en évidence de la granularité du contexte d'application de la tâche, et la mise au même niveau des éléments de comportement et des éléments contextuels. La visualisation proposée ici s'inspire de cette idée et automatise la construction d'un graphe.

4.1 Expérience

Afin de comparer les utilisations faites d'une interface par un groupe d'utilisateurs dans un hypermédia clos (c'est-à-dire où la navigation est contrainte à un espace défini), une expérience de consultation de 2 pages Internet sur un navigateur pour répondre à des questions a été menée. Le test est composé de 4 questions et s'accomplit en 6 sous-tâches (la première des sous-tâches est la consultation libre des documents, les suivantes sont l'affichage d'une question et la saisie de sa réponse, la dernière étant la fermeture de la page).

Les questions ont été étudiées pour faire appel à différents processus cognitifs. Certaines questions demandent des renseignements très précis sur le contenu d'une des deux pages, alors que d'autres demandent une réponse portant sur le contenu ou la forme générale des pages. Une trentaine d'utilisateurs de différents niveaux d'expertises ont effectué cette expérience.

4.2 Résultats

Cette représentation doit être manipulée avec précaution tant les modifications qui y ont été apportées pour faire apparaître les sous-tâches transforment la lecture. En effet, l'échelle temporelle (horizontal) est ici différente pour chacune des sous-tâches, et la comparaison entre les traces ne doit pas se faire par une analyse des rythmes ou de la rapidité dépendant du temps. En conséquence, l'aspect séquentiel des actions réalisées par un utilisateur est seul conservé, et on pourra analyser le rythme d'une action sur une sous-tâche en le comparant avec le rythme des actions réalisées dans la même sous-tâche par le même utilisateur.

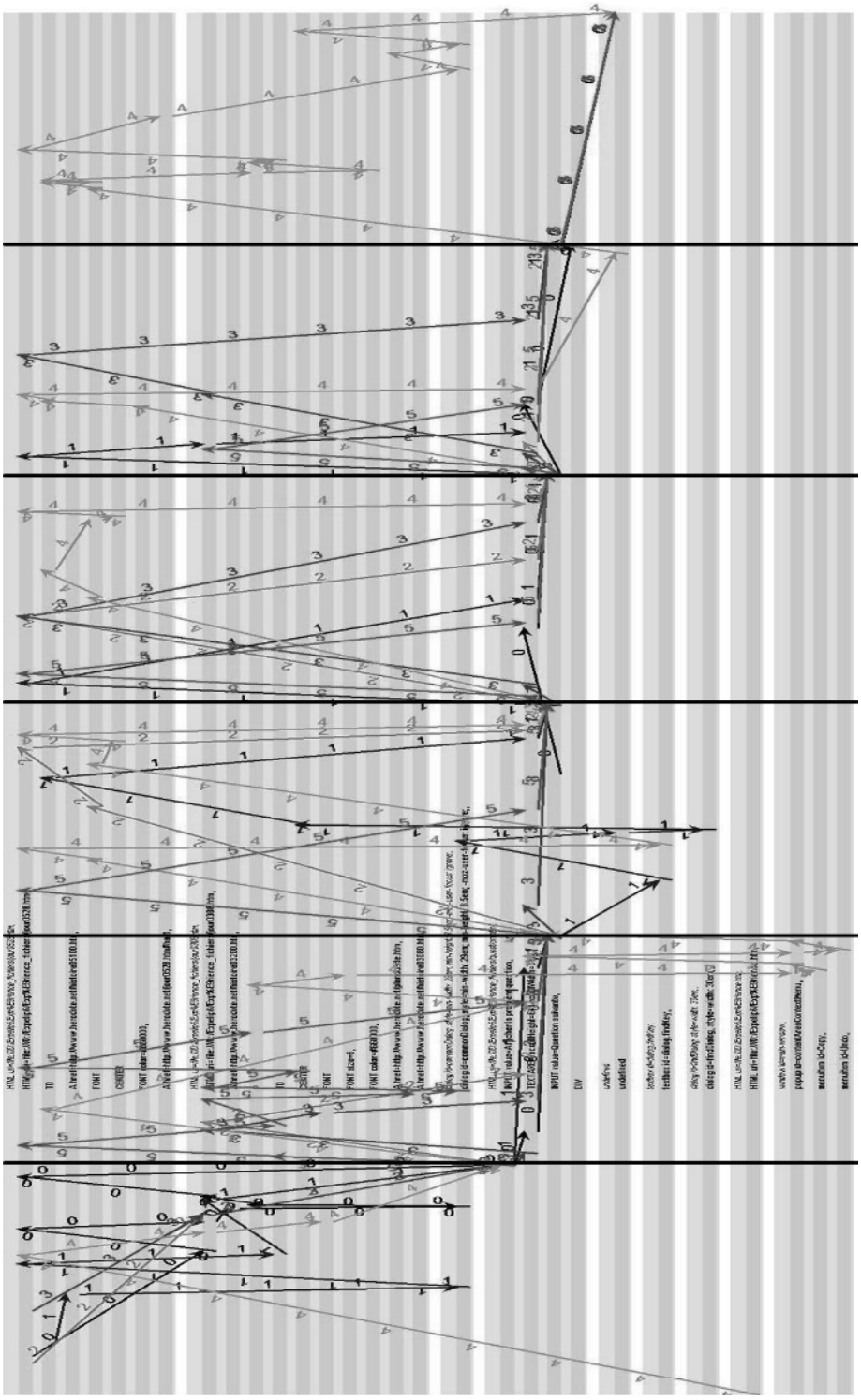


Figure 3. Visualisation des traces de 5 utilisateurs avec représentation des sous-tâches

La visualisation proposée incorporant la visualisation des sous-tâches est présentée dans la Figure 3, chaque sous-tâche étant délimitée par les traits verticaux.

Cette figure représentant les traces de 6 utilisateurs fait apparaître plusieurs propriétés. L'utilisateur 5 n'a pas consulté les documents avant l'affichage de la première question. L'utilisateur 4 est le seul qui a continué à lire les documents après avoir répondu à toutes les questions. Cet utilisateur semble également expert dans la manipulation de l'interface, puisqu'il a utilisé des outils, les barres grises horizontales du bas de la figure qui représente les fonctions de copier/coller ($2^{ième}$ étape) et de recherche ($3^{ième}$ étape) dans les menus du navigateur internet. L'utilisation de cette dernière fonction n'a d'ailleurs pas été fructueuse puisqu'elle est suivie d'une deuxième consultation de la page Internet où la réponse a finalement été trouvée.

Cette représentation peut être comparée à celle des graphes contextuels car elle met en parallèle l'accomplissement d'actions effectuées par différents utilisateurs dans un même contexte. Il s'agit pour tous les utilisateurs de répondre à des questions à partir des mêmes documents en ayant chacun accompli les sous-tâches précédentes mais de façon différentes. Cette différence constitue donc le contexte de réflexion de chaque utilisateur et influe sur l'action portée dans la tâche courante.

5 Conclusion

Une visualisation de parcours d'un utilisateur sur une interface graphique automatisée par les évènements que génère cette interface est proposée ici. Deux exemples d'analyse cognitive de parcours proposent des interprétations de cette visualisation. Les résultats ont permis de comparer l'utilisation de deux types d'interface et de concevoir un nouvel outil pour la pédagogie sur manuel scolaire numérique. Une interprétation propose une remise en contexte de l'action par une description séquentielle de la tâche.

D'autres outils permettant d'autres types d'interprétation peuvent être ajoutés. La lecture d'un plus grand nombre de traces et une représentation non contraignante pour les traces longues ou les interfaces complexes doivent encore être facilitée.

Références

Brézillon P., Tijus C. *Une représentation basée sur le contexte des utilisateurs à travers leurs pratiques*. Atelier sur la modélisation utilisateurs et personnalisation de l'interaction homme-machine. EGC 2005. Paris.

Damez M. *Méthode de récolte de traces de navigation sur interface graphique et visualisation de parcours*. Article pour démonstration de logiciel, EGC 2006. Paris. (à paraître)

Delort J.Y., Bouchon-Meunier B., Rifqi M. *VISS : A Tool for Visualizing Clues about the Users' Information Needs and Their Information-Seeking Tactics*. Poster Proceedings of The Fourteenth International ACM Conference on Hypertext and Hypermedia. Nottingham, United Kingdom. 2003.

John B., Vera A., Matessa M., Freed M., Remington R. *Automating CPM-GOMS*. SIGCHI Conference on Human factors in computing systems, Minneapolis, USA. 2002.

Pitkow J., Bharat K. *Webviz: A tool for world wide web access log analysis.* Advance Proceedings First International World-Wide Web Conference, p. 217-277, May 1994. Geneva, Switzerland,

Rumbaugh J., Jacobsen I., Booch G. *UML 2.0.* Campus Press. Décembre 2004.

Spiliopoulou M. et Faulstich L.C. WUM : a Web Utilization Miner. Workshop on the Web and Data Bases (WebDB98), p. 109-115, 1998. Valencia, Spain. March 1998.

Vignaux G. *Les relations sémantiques et cognitives appliquées aux notions dans un texte*. http://www.colisciences.net/modules.php?name=accueil&pa=showpage&pid=11, 2002

Vers une méthodologie rigoureuse de conception des langages graphiques s'appuyant sur les sciences cognitives

Jean-Baptiste Lamy*, Catherine Duclos*
Vincent Rialle**, Alain Venot*

*LIM&BIO, UFR SMBH Université Paris 13
74 rue Marcel Cachin, 93017 Bobigny cédex
jibalamy@free.fr,
http://www.limbio-paris13.org/
**TIMC-IMAG, Université de Grenoble 1, faculté de médecine
domaine de la Merci, La Tronche, France

Résumé. Le volume d'information et de connaissance ne cesse de croître. Les langages graphiques, parce qu'ils peuvent être lus plus rapidement que le texte, sont une solution à ce problème. Cependant, la conception de ces langages suit rarement une méthodologie rigoureuse. À partir d'éléments issus des sciences cognitives comme la sémiologie graphique et la théorie de la Gestalt, nous proposons des règles pour la conception de ces langages, notamment pour la représentation graphique de relation est-un.

1 Introduction

Dans de nombreux domaines, le volume d'information et de connaissance connaît une explosion sans précédent : en médecine par exemple, de plus en plus de données d'essais cliniques, de guides de recommandations, de connaissance sur les médicaments sont disponibles.

L'utilisation de langages graphiques est une solution qui a été proposée à ce problème, avec des succès comme l'étiquetage des produits chimiques, les panneaux routiers ou le langage de modélisation UML (*Unified Modeling Language*). Ces langages permettent de présenter graphiquement des informations ou des connaissances, et peuvent être lus plus rapidement que le langage naturel textuel. En revanche, ils imposent des besoins matériels plus importants (couleur,...) et n'ont pas la précision du texte.

Les langages graphiques peuvent être utilisés à différents niveaux :

– Lors de la fouille d'information : un langage graphique permet de représenter graphiquement les informations à fouiller. Par exemple, des langages graphiques très simples sont fréquemment utilisés pour représenter par des icônes des fichiers informatiques dans un gestionnaire de fichier : il est aisé de retrouver des fichiers informatiques présentant des caractéristiques données grace à des icônes spécifiques (figure 1, gauche). Les langages graphiques sont aussi utilisés dans

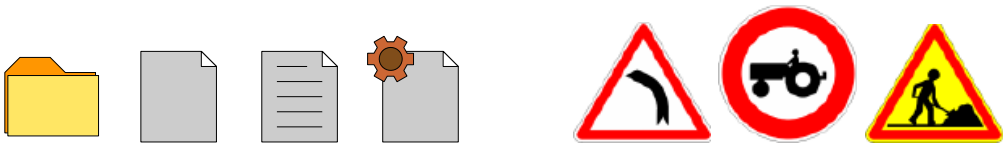


Fig. 1 – *Exemples de langages graphiques. A gauche : icônes utilisés dans un gestionnaire de fichier : répertoire, fichier, fichier texte, fichier exécutable. A droite : panneaux routiers : attention virage à gauche, interdit aux véhicules agricoles, travaux routiers.*

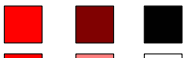
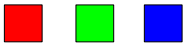
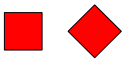
certaines techniques de visualisation d'information comme les glyphes (Chuah et Eick, 1997; Osawa, 2002; Erbacher, 2002).
- Pour présenter à l'utilisateur final les connaissances extraites : l'utilisation d'un langage graphique est particulièrement indiqué lorsque les connaissances doivent être lues très rapidement, ou lorsqu'elles sont très volumineuses. Un exemple type est celui des panneaux routiers dont la lecture doit être la plus rapide possible (figure 1, droite).

Cependant, la méthodologie de construction de ce langages graphiques est rarement décrite, et l'absence de méthodologie rigoureuse peut conduire à des langages peu cohérents, difficiles ou inadaptés aux besoins. Nous considérons que la conception d'un langage graphique devrait prendre en compte deux facteurs : la nature des informations à représenter, et les capacités de la vision humaine. Dans ce papier, nous nous intéressons au second facteur. Nous commencerons par rappeler deux théories issues du domaine des sciences cognitives : la sémiologie graphique de J. Bertin et la théorie de la *Gestalt*, puis nous nous appuierons dessus pour proposer des règles à suivre lors de la conception d'un langage graphique. Ces règles ont été appliquées à la conception d'un langage graphique pour les connaissances sur le médicament (contre-indications, effets indésirables, précautions d'emploi,...) au laboratoire LIM&BIO (Lamy et al., 2005).

2 État de l'art

2.1 La sémiologie graphique

La sémiologie graphique de Bertin (1999) est un modèle empirique qui a ensuite été partiellement validé par les modèles neurophysiologiques. Une image peut être décomposée en plusieurs éléments, que nous appellerons *signes*, chaque signe étant défini par plusieurs *variables rétiniennes* qui ont chacune des propriétés différentes (tableau 1) :
- Une variable est *ordonnée* si les différentes valeurs de la variable ont un ordre évident. Par exemple la couleur est ordonnée si l'on se fixe une échelle de couleur.
- Une variable *quantitative* peut exprimer des rapports numériques. Par exemple la taille est quantitative (un objet peut être deux fois plus grand qu'un autre).
- Le *nombre de valeurs* est le nombre maximum de valeurs différentes que l'oeil humain peut distinguer pour cette la variable. Par exemple, nous pouvons distinguer trois niveaux de saturation différents.

variable rétinienne	exemple	ordonnée	quantitative	nb de valeurs
taille		+	+	5
valeur		+	(+)	3
saturation		+	(+)	3
grain		+		3
couleur		(+)		10
orientation		(+)		12
forme, pictogramme		(+)	(+)	quasi-infinie

Tab. 1 – *Les variables rétiniennes et leurs propriétés d'après J. Bertin. "+" indique que la propriété est présente pour une variable, "(+)" indique qu'elle est présente sous certaines conditions (changement d'échelle,...).*

Lorsque l'on recherche visuellement un groupe de signes ayant des propriétés communes (par exemple "forme = carré"), deux situations peuvent se présenter :

- Il faut passer en revue chacun des signes un par un, et le temps nécessaire est proportionnel au nombre de signes. On parle alors de *perception attentive*.
- Le groupe de signes semble "surgir" hors de l'image, et peut être distingué très rapidement (moins de 200 ms, indépendamment du nombre de signes). On parle alors de *perception pré-attentive*.

Bertin ainsi que d'autres auteurs (Pylyshyn, 1994) ont montré que la perception pré-attentive n'est possible que si la recherche porte sur une seule variable rétinienne.

La sémiologie graphique a cependant ses limites : il n'est pas toujours possible de considérer les variables rétiniennes comme indépendantes les unes des autres, et certaines variables (notamment la couleur) se prêtent mieux à la perception pré-attentive.

2.2 La théorie de la Gestalt

La psychologie de la forme ou théorie de la *Gestalt* (Paul, 1979) a été fondée au début des années 20 par des psychologues allemands ; en allemand, *Gestalt* signifie forme mais aussi structure, organisation. Cette théorie générale a été appliquée à la perception visuelle : pourquoi dans une image donnée percevons-nous telle forme plutôt que telle autre ? Ce que nous percevons dans une image est la "meilleure" forme, c'est à dire la plus simple. Les lois de la ségrégation permettent de déterminer la simplicité d'une forme. Une forme simple :

- regroupe des éléments proches (loi de proximité),
- regroupe des éléments similaires (même couleur,...) (loi de ressemblance),
- est symétrique (loi de symétrie),
- correspond à une forme très connue (par exemple un cercle, un carré,...), même si la forme n'est pas complète (loi de clôture),

– les règles répétées plusieurs fois voient leur effet renforcé (loi de répétition).

3 Conception d'un langage graphique

La première étape de la conception d'un langage graphique consiste à analyser les informations ou connaissances que l'on souhaite représenter. Lorsque des textes sont disponibles, cette analyse peut se faire par des outils de Traitement Automatique du Langage (TAL). L'analyse doit en particulier déterminer les attributs qui définissent les termes à représenter, et les relations d'héritage "est-un" entre ces termes. Par exemple pour représenter des maladies, les attributs pourraient être la localisation anatomique (cardiaque, rénale,...), le type de maladie (cancer, infection,...) et la sévérité. Il existe aussi un grand nombre de relations est-un, comme "trouble du rythme" est une "maladie cardiaque".

Ensuite, il faut associer à chaque attribut une variable rétinienne pour le représenter, par exemple pour les maladies il est possible d'associer l'attribut localisation anatomique à la variable forme/pictogramme, et l'attribut sévérité à la variable taille. Enfin, les relations est-un entre les différents termes représentés doivent être explicitées. Les aspects cognitifs étudiés à la section précédente doivent être pris en compte dans ces deux dernières étapes.

3.1 Choix des variables rétiniennes

Lorsque l'on construit une représentation graphique, chaque attribut à représenter est associé à une variable rétinienne. La variable doit être choisie en tenant compte des propriétés de l'attribut et des propriétés de la variable. Par exemple, si l'on souhaite représenter la fréquence d'un effet indésirable définie par cinq classes : exceptionnel, rare, moyennement fréquent, fréquent et constant, il faudra une variable ordonnée (car les cinq classes sont elles-mêmes ordonnées) et capable de représenter au moins cinq valeurs (une par classe). Par conséquent, seules les variables taille, couleur et forme sont appropriées pour représenter ce concept.

En regardant le tableau 1, on constate que certaines variables sont plus intéressantes que d'autres, en particulier : la forme (nombre de valeur quasi-infini), la couleur (nombre de valeur assez élevé, et très discriminant visuellement) et la taille (ordonnée et quantitative). Les autres variables ne sont que rarement utilisées.

Enfin, il faut faire en sorte que les principales recherches effectuées par les utilisateurs puissent avoir lieu de manière pré-attentive, c'est à dire que ces recherches ne mettent en jeu qu'une seule variable rétinienne.

3.2 Représentation des relations est-un

Les relations est-un sont très fréquentes ; par exemple "trouble du rythme" est une "maladie cardiaque". Les langages graphiques doivent rendre explicites ces relations, de sorte qu'un utilisateur qui recherche visuellement les maladies cardiaques trouve aussi les troubles du rythme. Il faut donc que dans le signe qui représente "trouble du rythme", l'utilisateur puisse voir le signe de "maladie cardiaque", plus un signe spécifique

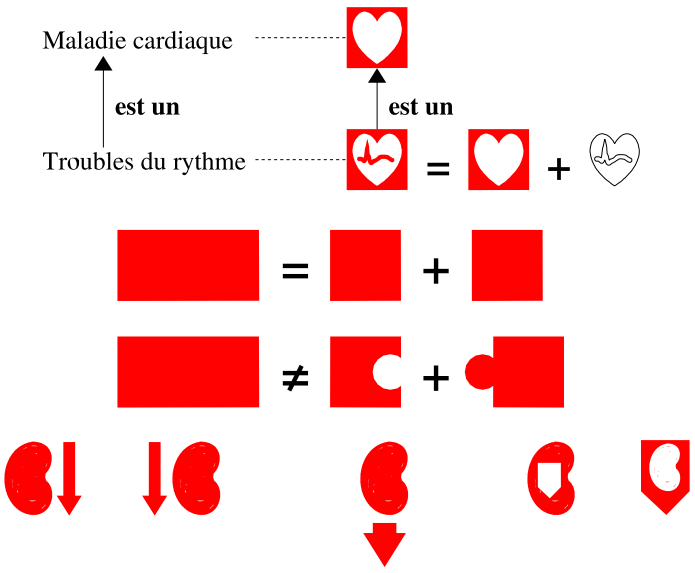


Fig. 2 – *En haut : exemple de représentation de relation est-un. Au milieu : le rectangle rouge est la combinaison de deux carrés rouges car c'est la coupure la plus simple selon les lois de la* Gestalt *(le carré étant une forme simple, répété deux fois de façon symétrique), mais il n'est pas la combinaison de deux "pièces de puzzle", car il n'est pas possible de deviner l'existence des deux pièces dans le rectangle. En bas : utilisation de juxtaposition ou d'inclusion pour signifier "insuffisance rénale" en combinant une flèche vers le bas et un rein.*

à "trouble du rythme" (figure 2, haut). Par contre, "risque de maladie cardiaque" ne doit pas être représenté en ajoutant un signe spécifique au signe de "maladie cardiaque", car un risque de maladie cardiaque n'est pas une maladie cardiaque.

De manière général : si B est un A, alors le signe qui représente B doit être la combinaison entre le signe qui représente A et un nouveau signe. Inversement, si B n'est pas un A, alors le signe de B ne doit pas être une combinaisons comprenant le signe de A. Les lois de ségrégation de la *Gestalt* permettent de déterminer si un signe résulte de la combinaison de deux autres ou non : en effet il ne suffit pas de superposer deux signes pour obtenir une combinaison valable (figure 2, milieu).

La combinaison des deux signes peut se faire de différentes manières : soit en juxtaposant les deux signes l'un à côté de l'autre, soit en incluant l'un des signes dans l'autre (figure 2, bas). Dans la juxtaposition, les deux signes jouent des rôles équivalent, mais risquent d'être considérés comme indépendants l'un de l'autre, ou contraire d'interagir ensemble (sur la figure 2, bas, au centre, le rein semble le départ de la flèche). Dans l'inclusion, les deux signes jouent des rôles différents, et le signe placé à l'intérieur rend le signe extérieur plus difficile à lire. Il faut donc toujours placer le signe le plus simple à l'extérieur. Dans l'exemple figure 2, le rein est un organe et la flèche, qui signifie

insuffisance, est un type de maladie ; tout deux jouent donc des rôles différents, c'est pourquoi nous choisirons l'inclusion. La flèche étant plus simple que le rein, elle sera placé à l'extérieur. La meilleure représentation est donc la dernière.

4 Conclusion

Nous avons montré comment deux aspects des sciences cognitives, la sémiologie graphique et la théorie de la *Gestalt*, pouvaient être mis à profit lors de la conception d'un langage graphique, afin de choisir les bonnes variables rétiniennes pour représenter les bons attributs, et afin d'expliciter les relations est-un entre les termes représentés. Ces considérations pourraient servir de base à une méthode de conception rigoureuse pour les langages graphiques, qui intégrerait également les apports d'autres théories ou travaux.

Par ailleurs, rappelons qu'une méthode, aussi rigoureuse soit elle, ne dispense pas d'évaluer soigneusement les langages graphiques ainsi que les application les utilisant.

Références

Bertin, J. (1999). *Sémiologie graphique : Les diagrammes - Les réseaux - Les cartes*. Paris-La Haye : Editions de l'Ecole des Hautes Etudes en Sciences.

Chuah, M. et S. Eick (1997). Glyphs for software visualization. In *International Workshop on Program Comprehension*, Dearborn, MI, USA, pp. 183–191.

Erbacher, R. (2002). Glyph-based generic network visualization. In *Proceedings of the SPIE '2002 conference on Visualization and Data Analysis*, pp. 228–237.

Lamy, J.-B., C. Duclos, V. Rialle, et A. Venot (2005). Which graphical approaches should be used to represent medical knowledge ? In *Stud Health Technol Inform. (Proceedings of Medical Informatics Europe MIE2005)*, Volume 116, Geneva, pp. 719–724.

Osawa, N. (2002). Visualization of inheritance relationships using glyphs. *IEICE Trans. on Information and Systems E85-D*(1), 275–282.

Paul, G. (1979). *La psychologie de la forme*. Paris : Flammarion.

Pylyshyn, Z. (1994). Some primitive mechanisms of spatial attention. *Cognition* (50), 363–384.

Summary

The amount of information and knowledge is endlessly growing. Since they can be read faster than texte, graphical languages are a solution to this problem. However, the conception of these languages seldom follows a rigorous methodology. Relying on some elements from cognitive sciences like graphical semiology or the Gestalt theory, we propose some rules for conceiving such languages, in particular for representing graphically is-a relation.

Session 3 : Visualisation et analyse de données

Visualiser les distorsions dans les techniques de projection continues

Michaël Aupetit

Classification de distributions par décomposition de mélange de copules archimédiennes : choix de la dimension des copules par visualisation

Etienne Cuvelier, Monique Noirhomme-Fraiture

Visualiser les distorsions et reconstruire la topologie dans les techniques de projections continues

Michaël Aupetit

CEA – Dépt. Analyse Surveillance Environnement
BP12
91680 Bruyères-le-Châtel
michael.aupetit at cea.fr

Résumé. Les techniques de projections continues (Analyse en Composantes Principales, Mapping Non-Linéaire...) permettent d'analyser visuellement des données de grande dimension. Nous proposons deux techniques permettant de visualiser dans l'espace de projection les zones de compressions, étirements, recollements et déchirements de variétés. Cela permet de discriminer des artefacts de projection, les aspects caractéristiques des données originales, et de reconstituer en partie la topologie des variétés support de ces données.

1 Introduction

1.1 Analyse exploratoire par projection de données

L'analyse exploratoire d'un ensemble de données multi-dimensionnelles est essentielle pour permettre à l'analyste d'appréhender et de traiter des bases de données toujours plus importantes. Nous considérons le cas de données vectorielles d'un espace euclidien $E=R^D$ appelé "espace d'origine" par la suite. Parmi d'autres techniques (Aupetit, 2005), la projection de ces données en 2 dimensions dans $F=R^2$ euclidien, permet de saisir en partie leur topologie et leur géométrie, à condition que la projection soit fidèle.

Nous nous intéressons aux projections continues linéaires telles que la projection parallèle aux axes, l'Analyse en Composante Principales (ACP) (Jollife, 1986) ou le Scaling Multi-Dimensionnel (Torgerson, 1952), et non-linéaires telles que le Mapping Non-Linéaire (Sammon, 1969) et l'Analyse en Composantes Curvilignes (ACC) (Demartines et Hérault, 1997).

1.2 Visualiser les distorsions

Venna et Kaski (2001) soulignent l'importance pour l'analyste de savoir si les points voisins de l'espace de projection correspondent à des données voisines dans l'espace d'origine, car c'est sur ces proximités que sont fondées les analyses visuelles. Ces auteurs

proposent une mesure globale (un nombre) pour caractériser cette préservation de voisinage. Nous proposons ici, deux techniques de visualisation de ce type d'information basées sur la coloration des cellules de Voronoï associées aux projections ou aux paires de projections voisines.

2 Notations

Nous considérons une matrice (N,N) de distances X obtenue par le calcul de toutes les distances X_{ij} entre paires de points $\{x_i,x_j\}$ de $\underline{x}=(x_1,\ldots,x_N)$ dans l'espace d'origine E, et l'ensemble des projections correspondantes $\underline{y}=(y_1,\ldots,y_N)$ dans l'espace F euclidien avec la matrice de distance Y associée.

La cellule de Voronoï V_i associée au point y_i est définie par (Okabe et al., 1992):

$$\forall y_i \in \underline{y}, V_i = \left\{ v \in F \middle| \forall y_j \in \underline{y}, (v-y_i)^2 \leq (v-y_j)^2 \right\}$$

Nous considérons aussi les paires de projections pour lesquelles les cellules de Voronoï sont adjacentes, donc les liens reliant ces paires $\underline{L} = \left\{ \{i,j\} \in (1,\ldots,N)^2 \middle| i \neq j, V_i \cap V_j \neq \phi \right\}$ correspondent aux liens du graphe de Delaunay des projections $\underline{y}$ dans l'espace F (Okabe et al., 1992). La cellule de Voronoï du segment de droite joignant y_i et y_j, est définie par:

$$\forall \{i,j\} \in \underline{L}, V_{ij} = \left\{ v \in F \middle| \forall \{k,l\} \in \underline{L}, d_{ij}(v) \leq d_{kl}(v) \right\}$$

avec $d_{ij}(v) = \min_{\alpha \in [0,1]} \left(v - \left(\alpha y_i + (1-\alpha) y_j \right) \right)^2$.

Ces cellules de Voronoï couvrent l'espace F et permettent un association visuelle rapide entre la mesure (couleur) et le point ou couple de points auquel elle se rapporte.

3 Mesures de distorsions

Nous proposons une mesure de "proximité" $prox_{i|s}$ associée à un point de référence y_s sélectionné dans F par l'analyste. Nous affichons par un niveau de gris dans chaque cellule de Voronoï V_i, la distance X_{is} entre les pré-images correspondantes x_i et x_s dans E (clair = proche, foncé = éloigné) :

$$prox_{i|s} = 1 - \frac{X_{is}}{\max_k (X_{ks})}$$

Cette mesure permet de détecter les recollements et déchirements de variétés au point de référence y_s (distorsions topologiques). Un recollement a lieu si deux points très éloignés et qui ne sont pas voisins dans l'espace d'origine, se retrouvent voisins et proches dans l'espace de projection (cellules de Voronoï adjacentes mais de niveaux de gris très différents). Un déchirement a lieu si deux points voisins et proches dans l'espace d'origine se retrouvent séparés par d'autres points dans l'espace de projection (cellules de Voronoï non adjacentes de couleur claires, séparées par des cellules plus foncées). Cette mesure ne donne qu'une vue parmi les N possibles liées au choix de y_s.

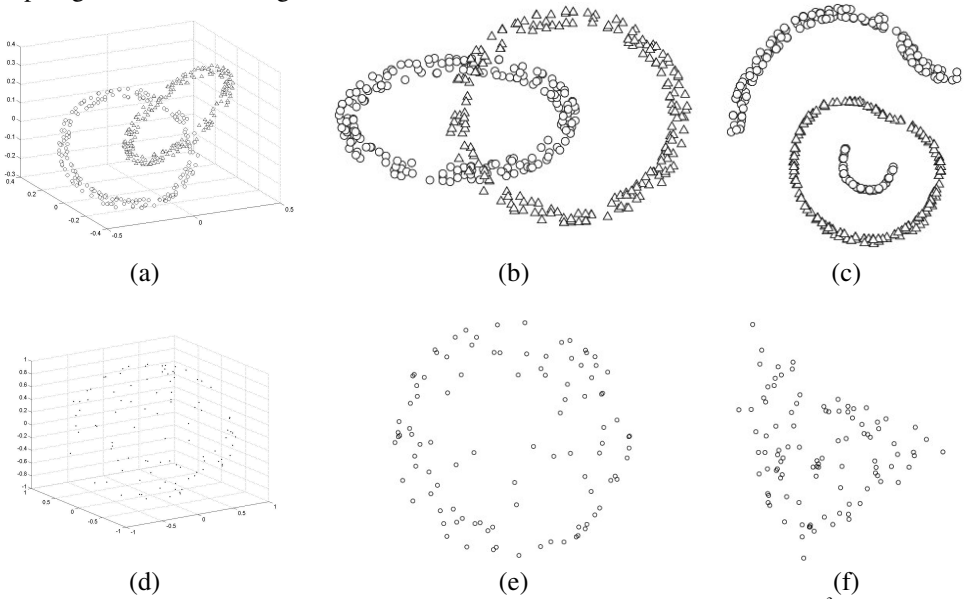
Nous proposons alors une mesure de distorsion $distor_{ij}$ associée à une paire de projections $\{i, j\} \in \underline{L}$. On mesure cette fois-ci l'écart entre X_{ij} et Y_{ij} et on l'affiche par un code de couleur: noir = étirement ($X_{ij}<Y_{ij}$), gris = aucune distorsion ($X_{ij}=Y_{ij}$) et blanc = compression ($X_{ij}>Y_{ij}$):

$$\forall \{i, j\} \in \underline{L}, distor_{ij} = \frac{X_{ij} - Y_{ij}}{\max_{\{k,l\} \in \underline{L}} \left(\left| X_{kl} - Y_{kl} \right| \right)}$$

Cette mesure permet de détecter les zones de compressions et d'étirements (distorsions géométriques) ainsi que les recollements de variétés. Elle donne une vue d'ensemble de ces distorsions mais elle ne permet pas cependant de détecter les zones de déchirement.

4 Expériences

Nous analysons des données issues de deux anneaux entrelacés et d'une sphère de R^3 dont nous connaissons la géométrie et la topologie. Ces données sont projetées par ACP (Jollife, 1986) et ACC (Demartines et Hérault, 1997). Nous présenterons une application à des données réelles lors de l'atelier. La figure 1 présente les données originales et leurs projections telles qu'elles apparaissent habituellement (Les marques permettent de différencier les données provenant de chaque anneau). Il faut bien sûr imaginer que nous ne connaissons des données que leur projection 2D sans marques. Que peut-on déduire de ces projections? En quoi les caractéristiques que l'on observe sont-elles représentatives des données originales ? Nous montrons comment les techniques de visualisation proposées permettent d'évaluer la qualité des projections et même de retrouver en partie la géométrie et la topologie des données originales.



(a) (b) (c)

(d) (e) (f)

FIG. 1 - *Echantillon d'un anneau entrelacé (a) et d'une sphère (d) de R^3. Projections correspondantes par ACP en (b) et (e), et par ACC en (c) et (f).*

4.1 ACP de deux anneaux entrelacés

4.1.1 Mesure de proximité

Un point de référence est sélectionné au niveau du croisement des deux anneaux dans F (Figure 2a). Un zoom sur cette région (Figure 2b) montre un recollement de variétés: tous les point associés aux cellules de Voronoï de couleur claire sont proches du point sélectionné (cellule blanche) dans l'espace d'origine E, tandis que les autres sont loin dans E, bien que tous soient projetés proches du point sélectionné dans F. Le même phénomène serait visible au croisement du bas, mais pas le long de chaque anneau ce qui permet d'en déduire que le nuage de point d'origine est constitué de deux composantes connexes distinctes (les deux anneaux). C'est un indice sur la topologie des données dans l'espace d'origine.

Tous les points de l'anneau de droite (Figure 2c) sont équidistants de la pré-image du point de référence dans l'espace d'origine (même couleur grise), ces points sont donc tous sur une sphère centrée en ce point dans E. C'est un indice sur la géométrie des données dans E.

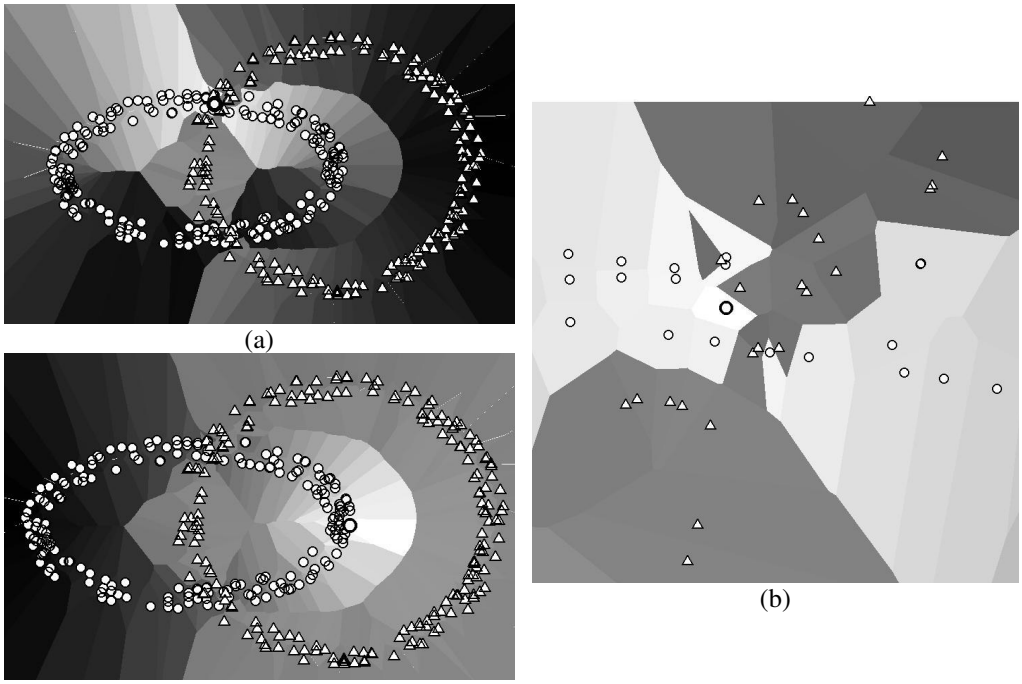


FIG. 2 – *Recollement de variétés détecté en (a) (zoom en (b)). Indice géométrique en (c).*

4.1.2 Mesure de distorsion

On visualise par la couleur des cellules de Voronoï des segments du graphe de Delaunay des projections, si les extrémités de ces segments ce sont rapprochées ou éloignées durant la

projection (Figure 3). En (a), le noir représente l'étirement maximal et le blanc la compression maximale. Cette mesure montre que les distances entre les points situés de part et d'autre d'un même anneau sont compressées indiquant pour l'anneau de droite que si c'est une ellipse à l'origine, elle n'est pas autant écrasée (indice géométrique). Par contre, le long des anneaux les distances sont bien préservées (couleur grise) sauf au niveau des deux croisements (zoom en (b)) où apparaissent de fortes compressions confirmant le recollement de deux variétés initialement disjointes.

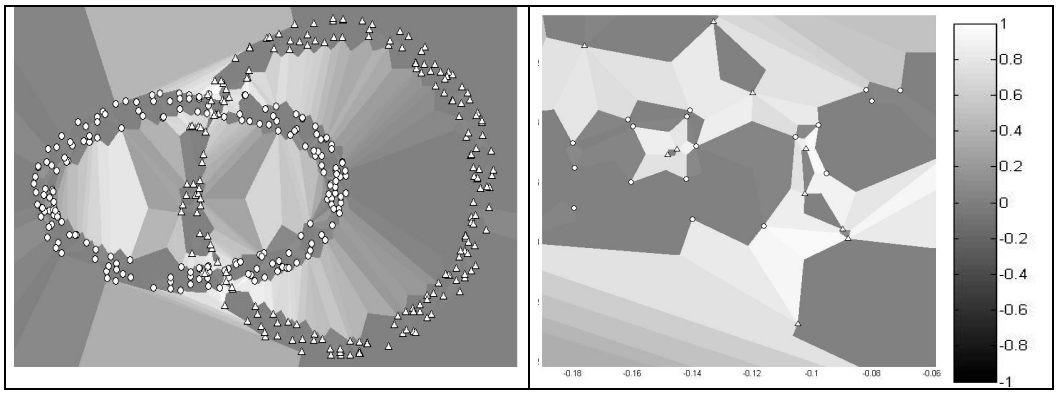


FIG. 3 – (a) mesure de distorsion indiquant une bonne préservation des distances le long des anneaux, une compression de l'anneau de gauche et un recollement de variétés au niveau des croisements des deux anneaux (Zoom du croisement inférieur en (b)).

4.2 ACC de deux anneaux entrelacés

Sur la figure 4a, la mesure de distorsion montre de fortes compressions (blanc) et quelques forts étirements (noir), mais ces éléments ne nous permettent pas de retrouver la topologie originale des données. Nous avons cependant un indice que les 3 composantes connexes observées dans F ne sont pas plus de 3 morceaux initiaux qui auraient été artificiellement recollés par la projection, car les distances le long de ces composantes sont bien préservées (couleur grise).

La mesure de proximité associée aux points extrêmes de la composante du haut montre sur les figures 4b et 4c, que cette composante et celle enclose dans l'anneau du bas sont connectées et comment elles le sont (les parties claires sont connexes dans l'espace d'origine). On détecte par cette mesure, un déchirement de variété. C'est un indice sur la topologie des données dans l'espace d'origine.

La mesure de proximité associée au point en blanc sur la figure 4c montre aussi que les deux morceaux artificiellement séparés par la projection, sont à la surface d'une sphère centrée sur ce point dans l'espace d'origine (même couleur grise pour tous ces points). C'est un indice sur la géométrie des données dans l'espace d'origine.

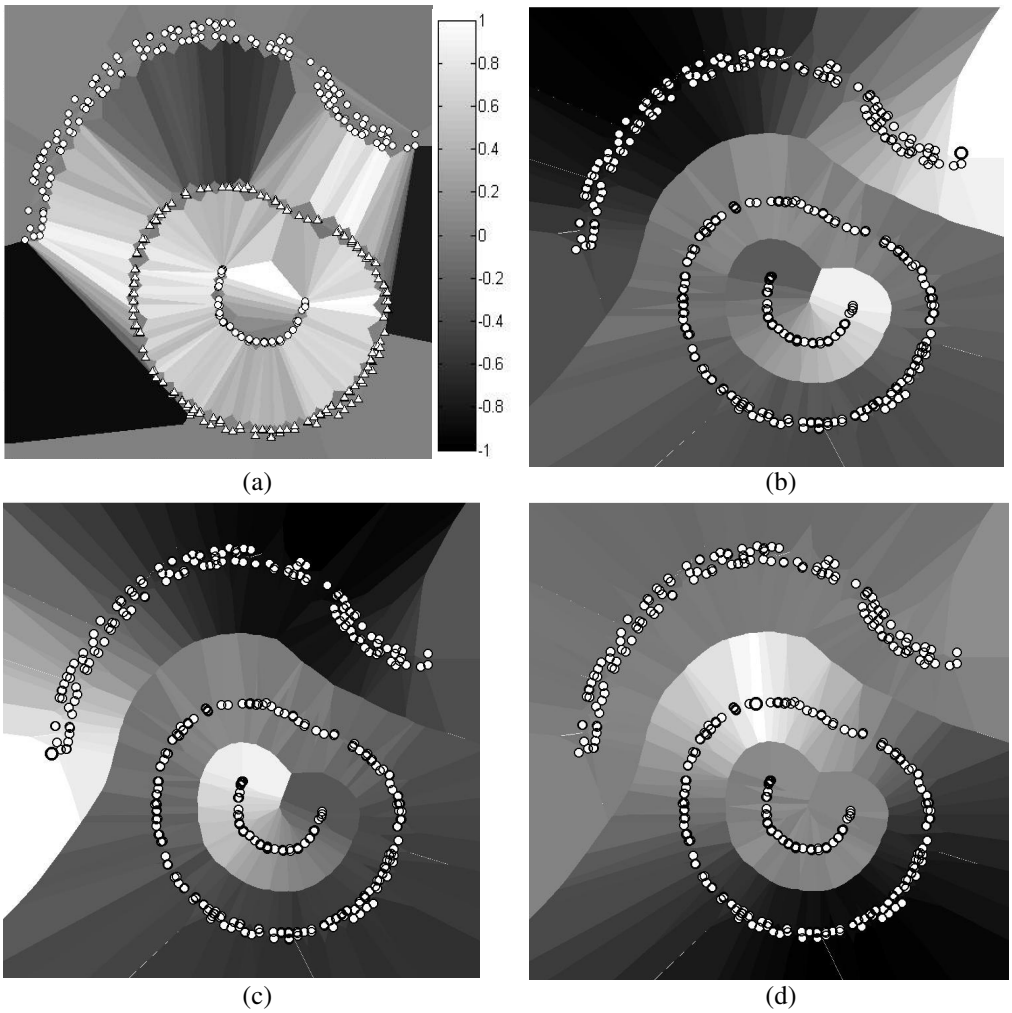
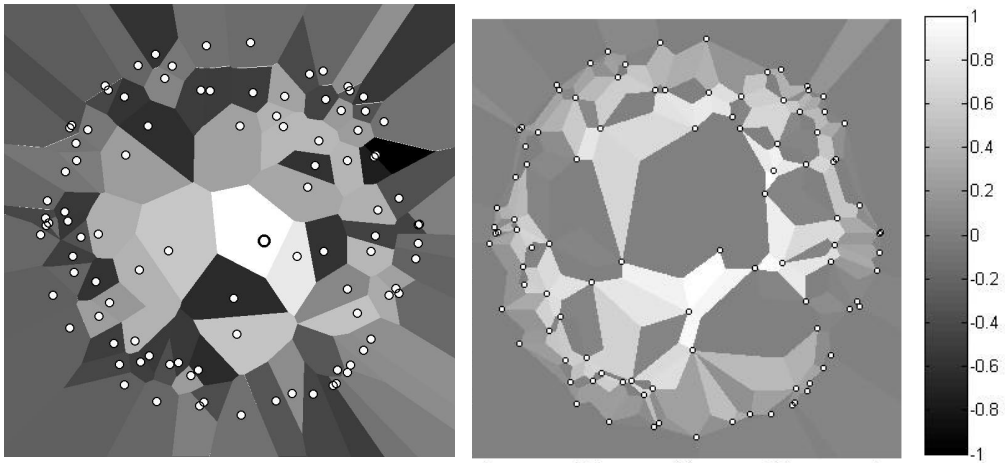


FIG. 4 – *(a) Mesure de distorsion confirmant la présence d'au plus 3 composantes connexes dans l'espace d'origine. (b-c) La mesure de proximité permet de reconstituer la topologie des variétés d'origine. (d) La mesure de proximité permet de découvrir un indice géométrique.*

4.3 ACP d'une sphère

La mesure de proximité montre sur la figure 5a, une alternance de niveaux de gris très différents sur des cellules voisines indiquant un recollement de variété : l'arrière et l'avant de la sphère sont projetés au même endroit.

La mesure de distorsion sur la figure 5b, montre le même phénomène: en gris les paires de points issues du même côté de la sphère (distances préservées), en blanc les paires initialement sur deux faces opposées (distances compressées). On note aussi une préservation des distances le long de l'enveloppe convexe des projections.



(a) (b)

FIG. 5 – *(a) Mesure de proximité indiquant un recollement de variétés. (b) Mesure de distorsion indiquant le même recollement.*

4.4 ACC d'une sphère

La mesure de proximité montre sur la figure 6a que les sommets de la forme triangulaire prise par les projections, sont en fait connectés dans l'espace d'origine. Un déplacement le long d'un côté du triangle (figure 6b) montre que ce côté est composé de deux moitiés jointes dans l'espace d'origine. On détecte ainsi un déchirement de la sphère que l'ACC a découpée selon trois lignes partant en étoile d'un même point, pour pouvoir l'aplanir en limitant les distorsions locales.

La figure 6c montre que l'on ne détecte ni déchirement, ni recollement au milieu du triangle, indiquant que la majorité des distorsions sont concentrées sur les côtés du triangle.

Sur la figure 6d, la mesure de distorsion ne permet pas de détecter le déchirement aussi clairement que la mesure de proximité, mais elle montre que des étirements ont lieu sur les côtés du triangle, et que les distances locales à l'intérieur du triangle sont bien préservées.

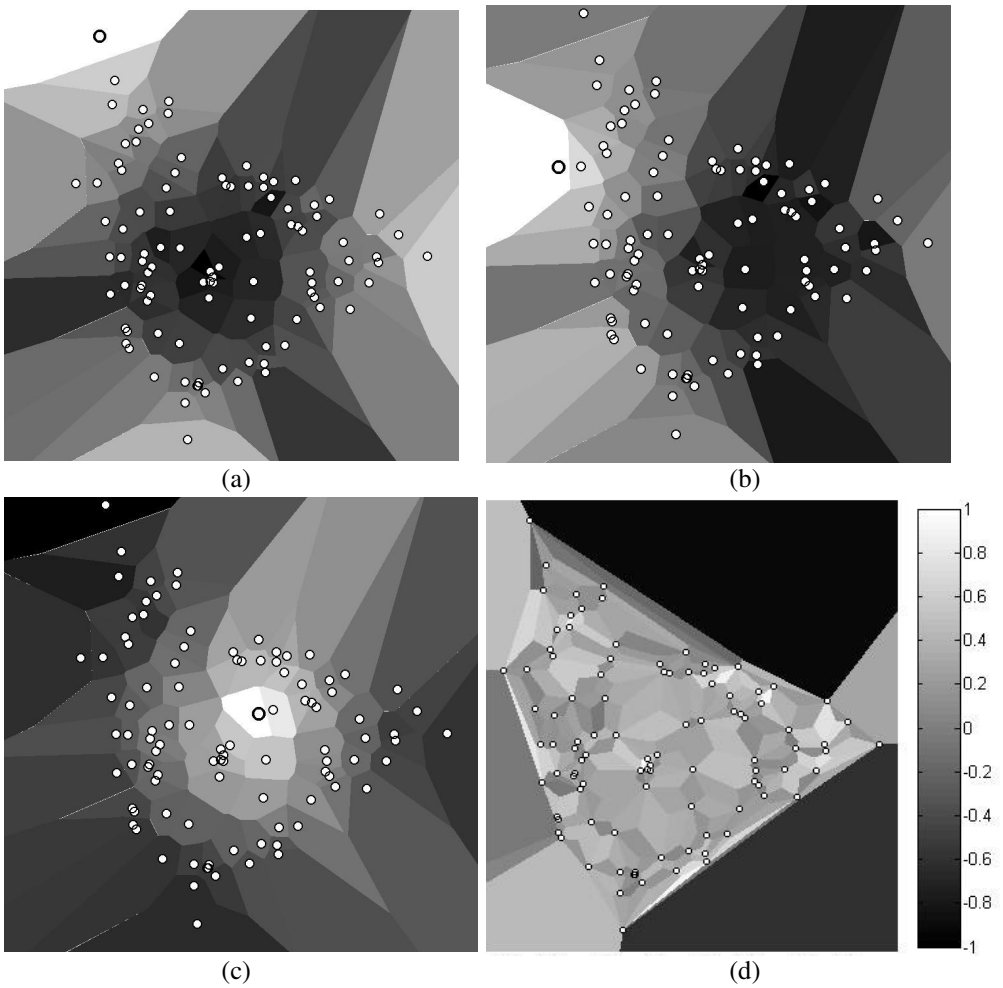


FIG. 6 – *La mesure de proximité permet de reconstituer la topologie originale des données (a-b-c). La mesure de distorsion indique une bonne préservation des distances à l'intérieur du triangle.*

5 Discussion

Les méthodes présentées permettent en fait une analyse visuelle des mesures de distances ou de dissimilarités déjà contenues dans les matrices X dans E et Y dans F, corrélée à la position des points projetés. Les matrices Xet Y sont généralement calculées par les techniques de projection. En cela, aucun calcul de distance supplémentaire n'est nécessaire (sinon, le calcul de ces matrices est en $o(DN^2)$). Le calcul des mesures de proximité et de distorsion est en O(N). Le calcul des cellules de Voronoï des projections est en o(N.log(N)) (Okabe et al. 1992) et en o(N) pour l'affichage.

Le graphe de Delaunay se calcule en O(N.log(N)) dans le plan et produit o(N) segments. Le calcul des cellules de Voronoï des segments du graphe de Delaunay est en o(N). Tous les calculs de graphe ou de cellules de Voronoï s'effectuent dans le plan. La complexité globale de ces méthodes de visualisation est donc en O(N.log(N)) puis en o(N) une fois les cellules construites.

Si cette complexité est trop importante, on peut envisager de la réduire en ne considérant la visualisation que des mesures associées à un sous-ensemble représentatif des points projetés, obtenu par tirage aléatoire par exemple, puis se concentrer sur une zone intéressante ainsi révélée où l'on considère la totalité des points de cette zone. Cette voie reste à explorer ainsi que la définition d'autres mesures porteuses d'informations complémentaires.

Pour la mesure de proximité, on ne visualise qu'un ensemble parmi N, de N mesures pour chaque point de référence sélectionné. Il faut donc un outil interactif de visualisation pour cette mesure, qui permette de sélectionner facilement une projection de référence au gré de l'analyse exploratoire. Au contraire, la mesure de distorsion est unique, elle ne dépend pas d'un point de référence particulier.

L'échelle de couleur (ici des niveaux de gris) a aussi son importance pour faire ressortir telles ou telles caractéristiques. Un outil interactif permettrait de modifier à volonté cette échelle de couleur pour explorer visuellement le contenu des matrice E et F.

6 Conclusion

Nous avons proposé deux méthodes de visualisation des distorsions dans les techniques de projection continues. Les cellules de Voronoï associées aux points et aux paires de points voisins sont coloriées en fonction de mesures de distorsions appropriées permettant de séparer les artefacts de projection, des caractéristiques saillantes topologiques ou géométriques dans les données d'origine. Ces techniques sont un complément indispensable à l'analyste pour éviter les erreurs d'interprétations et extraire davantage d'information à partir des projections.

Nous envisageons le développement de ces méthodes pour le cas de projections en 3 dimensions, et leur intégration dans un outil interactif de manipulation et d'exploration de données.

Références

Aupetit, M., T. Catz (2005). High-dimensional labeled data analysis with topology representing graphs. *Neurocomputing*, 63 :139-169.

Demartines, P., J. Hérault (1997). Curvilinear Component Analysis: a Self-Organising Neural Network for Non-Linear Mapping. *IEEE Trans. on Neural Networks*, 8(1):148-154.

Jollife, I.T. (1986). *Principal Component Analysis*. New York: Springer Verlag.

Okabe, A., B. Boots, K. Sugihara (1992). *Spatial tessellations: concepts and applications of Voronoï diagrams*. Chichester: John Wiley.

Sammon, J.W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. On Computers*, C-18:401-409.

Torgerson, W.S. (1952). Multidimensional scaling I - Theory and methods. *Psychometrica*, 17:401-419.

Venna, J., S. Kaski (2001). Neighborhood preservation in nonlinear projection methods: an experimental study. Lecture Notes in Computer Science, 2130:485-491.

Summary

Continuous projection techniques such as the Principal Component Analysis and the Non Linear Mapping, allow to perform visual analysis of high-dimensional data. We propose two visualization techniques allowing to detect straight in the projection space, where compression, stretching, gluing and tearing of manifolds occured. This allows to separate between projection artifacts and salient features of the original data, and to recover partly their original topology.

Classification de distributions par décomposition de mélange de copules archimédiennes: choix de la dimension des copules par visualisation

Etienne Cuvelier*, Monique Noirhomme-Fraiture*

*Institut d'Informatique, Facultés Universitaires Notre-Dame de la Paix,
21 rue Grandgagnage, 5000 Namur, Belgique
etienne.cuvelier@info.fundp.ac.be, monique.noirhomme@info.fundp.ac.be

Résumé. En analyse symbolique, un objet complexe peut être décrit par une variable s'exprimant comme une distribution de probabilité. La classification d'un ensemble d'objets symboliques décrits par ce type de variable, peut être obtenue en appliquant une décomposition de mélange de copules archimédiennes sur les valeurs des distributions calculées en un nombre q de points distincts, appelés coupures. Jusqu'à présent ces coupures ont été choisies arbitrairement.
Dans cet article nous montrons d'abord de façon empirique sur quelques exemples que le taux d'erreur de classement varie avec le nombre de coupures et leur position. Nous proposons ensuite de fixer ces deux paramètres grâce à une interaction visuelle.

1 Introduction

En analyse symbolique (Bock et Diday, 2000) une variable peut, entre autre, être décrite par une distribution de probabilité continue. Dans ce cas, la classification en K groupes de N objets symboliques décrits par cette variable, peut être obtenue en appliquant une décomposition de mélange aux valeurs obtenues par l'échantillonnage des distributions calculées en q endroits distincts, appelés coupures. L'estimation des composantes du mélange est réalisée à l'aide de copules archimédiennes. Cette appproche a déjà été utilisée avec succès par (Vrac et al., 2001) sur des données atmosphériques avec deux coupures. Nous étendons cette approche en permettant d'utiliser un nombre plus grand de coupures (Cuvelier et Noirhomme-Fraiture, 2005).
Dans tout les cas, le choix des coupures se révèle déterminant pour la qualité de la classification, mais actuellement aucun critère de décision automatique n'existe pour effectuer ce choix. Nous proposons donc, à l'instar de (Poulet, 2003) et de (Guo, 2003), une coopération entre une technique automatique de classification et technique visuelle de choix des coupures. Ce choix des coupures par visualisation étant basé sur des heuristiques inspirées des résultats de simulations.
Dans le paragraphe 2 nous rappellerons brièvement l'algorithme des nuées dynamiques appliqué aux distributions de probabilité, et nous introduirons la notion de copule. Le paragraphe 3 traitera de l'influence du nombre et du choix des coupures dans la qualité de la classification. Nous détaillerons ensuite dans le paragraphe 4 l'intérêt du choix visuel des coupures.

2 Classification de distributions de probabilités

2.1 Distributions de distributions

Nous supposerons que nous avons comme base de travail un tableau T de n lignes et p colonnes, et que la j^{me} colonne contient des distributions de probabilités , c'est-à-dire que si nous notons Y^j la j^{eme} variable alors Y_i^j est une distribution $F_i(.)$ pour tout $i \in \{1, ..., n\}$. Dans ce qui suit nous noterons ω_i le concept décrit par l'objet symbolique de la i^{eme} ligne, et $F_{\omega_i}(.)$ la distribution associée. Pour effectuer la classification en K classes nous commençons par échantilloner les distributions en q valeurs $T_1, ..., T_q$, et donc pour chaque $i \in \{1, ..., n\}$ nous calculerons $F_i(T_1), ..., F_i(T_q)$. Si nous appelons Ω l'ensemble de tous les concepts, la distribution conjointe des valeurs $F_i(T_j)$ est définie par :

$$H_{T_1,...,T_q}(x_1, ..., x_q) = P\left(\omega \in \Omega : \{F_\omega(T_1) \leq x_1\} \cap ... \cap \{F_\omega(T_q) \leq x_q\}\right) \quad (1)$$

et est appelée distribution de distributions.
La méthode, classique, de décomposition de mélange consiste à considérer cette distribution comme étant la résultante d'un mélange de distributions :

$$H_{T_1,...,T_q}(x_1, ..., x_q) = \sum_{i=1}^{K} p_i \cdot H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i) \quad (2)$$

avec $\forall\, i \in \{1, ..., K\} : 0 < p_i < 1$ et $\sum_{i=1}^{K} p_i = 1$.
La distribution de la i$^{\text{eme}}$ classe étant donnée par $H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i)$, avec $\beta_i \in R^d$, et p_i étant la probabilité qu'un élément appartienne à cette classe.
La densité de chaque distribution est donnée par :

$$h(x_1, ..., x_q) = \frac{\partial^q}{\partial x_1 ... \partial x_q} H(x_1, ..., x_q) \quad (3)$$

2.2 Algorithme des nuées dynamiques

L'algorithme utilisé (Diday, 2002) est en fait une extension de la méthode des nuées dynamiques (Diday et al., 1974) dans le cas d'un mélange. L'idée principale est, alternativement, d'estimer au mieux la distribution de chaque classe, et ensuite de vérifier que chaque objet symbolique appartient à la classe de densité maximale. L'étape d'estimation est réalisée en maximisant un critère de qualité, ici la log-vraisemblance :

$$lvc(P, \beta) = \sum_{i}^{K} \sum_{\omega \in P_i} log(h(w)) \quad (4)$$

avec

$$h(w) = h_{T_1,...,T_q}(F_\omega(T_1), ..., F_\omega(T_q)) \quad (5)$$

La classification commence avec une partition initiale aléatoire, et les deux étapes suivantes sont donc répétées jusqu'à stabilisation de la partition :

- **Etape 1 : Estimation des paramètres**
  Déterminer le vecteur $(\beta_1, ..., \beta_K)$ qui maximise le critère de qualité.

– **Etape 2 : Distribution des objets symboliques dans les classes**
Les classes $(P_i)_{i=1,...,K}$, dont les paramètres ont été calculés à l'étape 1, sont construites comme suit

$$P_i = \{\omega : h(\omega, \beta_i) \geq h(\omega, \beta_m) \forall m\} \tag{6}$$

2.3 Copules

L'estimation de distributions multivariées et de leurs densités n'est pas toujours chose aisée, alors que l'estimation univariée pose moins de problèmes.
Ainsi dans notre cas, les marges des distributions $H^i_{T_1,...,T_q}(x_1, ..., x_q; \beta_i)$ définies par

$$G^i_{T_j}(x) = P\{\omega \in P_i : F_\omega(T_j) \leq x\} \tag{7}$$

peuvent être facilement estimées par

$$\widehat{G}^i_{T_j}(x) = \frac{card\{\omega \in P_i : F_\omega(T_j) \leq x\}}{card(P_i)} \tag{8}$$

et les densités associées par la méthode des noyaux (Silverman, 1986)

$$\widehat{g}_{T_i}(x) = \frac{1}{card(P_i) \cdot h} \sum_{\omega \in P_i} K\left(\frac{x - F_\omega(T_i)}{h}\right) \tag{9}$$

où K est une fonction noyaux (Gaussienne, Epanechnikov, Triangulaire,...) et h est le paramètre de lissage, qui peut être calculé à l'aide de l'Erreur Quadratique Moyenne Intégrée (Mean Integrated Square Error - MISE).
La notion de copule (Nelsen, 1999) permet d'utiliser ces estimations des marges pour reconstruire les distributions multivariées.
Par définition

$$C(u_1, ..., u_n) \textit{ est une copule}$$
$$\textit{ssi}$$
$$C \textit{ est une distribution multivariée dont toutes les marginales sont uniformes sur } [0, 1]$$

Les copules sont des outils précieux dans la modélisation des structures de dépendance grâce au théorème de Sklar :
Si $H(x_1, , x_n)$ est une distribution multivariée de marginales $F_1(x_1), ..., F_n(x_n)$
alors il existe une copule C telle que

$$H(x_1, ..., x_n) = C(F_1(x_1), ..., F_n(x_n)) \tag{10}$$

de plus, si $F_1, ..., F_n$ sont toutes continues, alors C est unique ; sinon C est unique seulement sur $domF_1 \times ... \times domF_n$ Les copules capturent la structure de dépendance et permettent de séparer la modélisation de celle-ci de la modélisation des marges.
Plusieurs techniques de construction des copules existent (Joe, 1997; Nelsen, 1999), et parmi celles-ci une méthode génère une classe importante de copules : les copules archimédiennes.

2.4 Copules Archimédiennes

Les copules archimédiennes sont définies par

$$C(\underline{u}) = C(u_1, ..., u_n) = \phi^{-1}\left(\sum_{i=1}^{n} \phi(u_i)\right) \tag{11}$$

où ϕ est une fonction, appelée générateur, de $[0, 1]$ vers $[0, \infty]$ telle que
- ϕ est une fonction continue strictement décroissante
- $\phi(0) = \infty$ et $\phi(1) = 0$
- ϕ^{-1} est complétement monotonique sur $[0, \infty[$ c-à-d que

$$(-1)^k \frac{d^k}{dt^k}\phi^{-1}(t) \geq 0 \tag{12}$$

quel que soit $t \in [0, \infty[$ et pour tout k.

Il existe quatre familles de copules archimédiennes bivariées qui possèdent une extension intéressante en dimension quelconque. Il s'agit des copules de Clayton, Gumbel, Frank et Joe. Dans le cadre de cet article nous n'utiliserons que la première : la copule de Clayton :

$$C_\theta(\underline{u}) = \left(\sum_{i=1}^{n}(u_i^{-\theta}) - n + 1\right)^{-\frac{1}{\theta}} \tag{13}$$

Une des propriétés des copules archimédiennes est que pour k fixé (avec $2 \leq k \leq n$) toutes les marges de dimension k d'une copule sont identiques. Ainsi les marges bidimensionnelles, obtenues à partir de l'expression (11) de la façon suivante,

$$\phi^{-1}\left(\phi(u_i) + \phi(u_j) + \sum_{k \neq i,j} \phi(1)\right) = \phi^{-1}\left(\phi(u_i) + \phi(u_j)\right) = C(u_i, u_j) \tag{14}$$

sont toutes modèlisées de la même façon, c'est-à-dire avec le même générateur ϕ.

3 Influence du nombre et du choix des coupures

Pour illustrer notre propos nous utiliserons un ensemble artificiel de données (cf. Figure 1). Cet ensemble est constitué de 4 groupes de distributions. En parcourant le graphe de gauche à droite, nous trouvons 45 distributions exponentielles (exp1), 45 distributions normales (norm1), 45 distributions exponentielles (exp2), et enfin 50 distributions beta (beta1).

Pour chaque distribution, 500 nombres aléatoires ont été générés suivant la loi choisie, ensuite on a estimé la distribution empirique.

Nous avons ensuite généré tous les ensembles de 2 à 7 coupures équidistantes (d'un multiple de 0.25) et ayant comme première coupure au moins -4, et comme dernière coupure au plus 4 :

$$\{\{T_k = d + k.s : k \in \{0, ..., q-1\}\} : q \in \{2, ..., 7\}; -4 \leq T_0; T_{q-1} \leq 4; s = k * 0.25\} \tag{15}$$

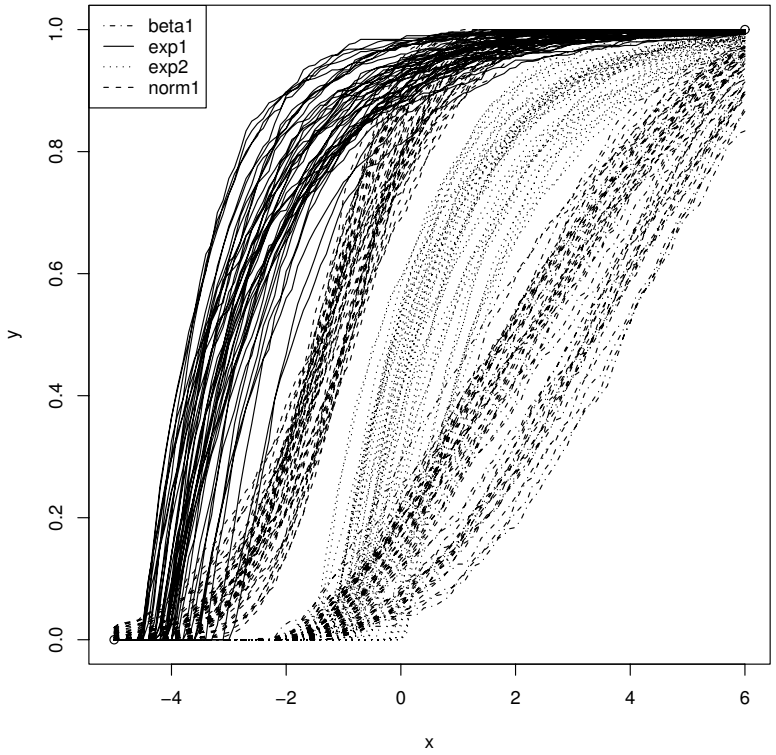


FIG. 1 – *Exemple d'ensemble de donnée*

Ensuite nous avons effectué les classifications sur base de chacun de ces ensembles de 2 à 7 coupures (820 ensembles). Ces classifications ont été réalisées en utilisant toujours la même partition initiale, et ce afin de pouvoir comparer les résultats des classifications.
Sur ces 820 ensembles de coupures, seuls 9 ensembles permettent de classer sans erreur les distributions. Ces ensembles sont les suivants : $\{-1.50, 0.50\}$, $\{-1.50, 1.00\}$, $\{-1.50, 1.25\}$, $\{-1.50, 1.50\}$, $\{-1.75, 1.50\}$, $\{-1.50, 2.00\}$, $\{-1.50, 2.25\}$, $\{-1.75, 2.25\}$, $\{-1.50, 2.50\}$.

Comme on peut le voir dans le tableau 1, si on choisit les coupures de manière arbitraire on peut obtenir des cas favorables et avoir un taux d'erreur proches de 0, mais aussi obtenir des cas très défavorables et avoir jusqu'à 51% d'erreur. Nous avons dès lors besoin d'heuristiques pour bien choisir les coupures.

Concentrons-nous maintenant sur les ensembles de coupures qui permettent d'effectuer la classification avec un taux d'erreur acceptable (10% maximum). En comparant les résultats par dimension des intervalles de la première et de la dernière coupure (Tableau 2) et le graphique nous pouvons observer les comportements suivants :

1. la longueur des intervalles des valeurs intéressantes pour les coupures varie de façon décroissante par rapport au nombre de coupures ;

q	Erreur moyenne	Erreur minimale	Erreur maximale
2	28,8%	0%	49,1%
3	33,1%	1,6%	51,3%
4	31,3%	4,3%	48,6%
5	28,5%	7%	48,6%
6	29%	5,4%	48,6%
7	28%	6,4%	49,1%

TAB. 1 – *Minima, maxima et moyennes par dimension*

q	Erreur moyenne	T_1	T_q
2	0.0334	[-2.50, -0.25]	[-1.75, 4.00]
3	0.0529	[-1.75, -0.75]	[0.25, 2.25]
4	0.0639	[-2.50, -0.50]	[0.25, 2.75]
5	0.0828	[-2.75, -1.75]	[-0.50, 1.25]
6	0.0567	[-2.75, -2.50]	[0.00, 2.25]

TAB. 2 – *Intervalles des premières et dernières coupures pour les meilleurs ensembles.*

2. l'augmentation du nombre de coupures n'améliore pas nécessairement le taux d'erreur ;

Le premier comportement est dû au fait que les copules archimédiennes modélisent de la même façon toutes les relations entres les marges (cf. supra). Cela permet de laisser "tomber" le début et la fin de l'ensemble des valeurs des distributions aux comportements différents de la partie centrale.
Le second comportement veille à maximiser le nombre d'informations par coupure compte tenu du nombre de coupures.

4 Intérêt du choix visuel

Les coupures ayant une influence directe sur la qualité des résultats, nous ne pouvons nous satisfaire d'un choix aléatoire. En nous inspirant des comportements ci-dessus nous pouvons émettre les heuristiques suivantes :

1. minimiser le nombre de coupures choisies ;
2. choisir des coupures qui maximisent le nombre de groupes discernables de valeurs le long de ces coupures, en veillant à ce que chaque groupe soit discerné un maximum de fois sur l'ensemble des coupures ;

Ainsi sur notre exemple, on peut visuellement repérer les zones suivantes dans l'espace possible des coupures :

- un faible intervalle situé autour de -0.5 où l'on peut presque distinguer 4 groupes distincts de valeurs ;
- un intervalle allant de -2.25 à 1.75 où l'on peut en chaque point distinguer au moins 3 groupes distincts de valeurs ;

– un intervalle allant de -3.25 à 4 où l'on peut en chaque point distinguer au moins deux distincts de valeurs ;

En suivant les heuristiques énoncées on peut par exemple choisir deux coupures, la première $-2 \leq T_1 \leq -1$ et la seconde $1 \leq T_2 \leq 1.75$. Dans nos tests les classifications qui ont utilisé des coupures respectant ces conditions ont les résultats suivants pour les taux d'erreurs : moyenne = 14,2%, min = 0% , max = 27%, ce qui représente une substantielle augmentation par rapport au choix aléatoire.

5 Conclusions et perspectives

Dans cet article nous avons montré comment le choix des coupures nécessaires à la classification pouvait se faire aisément visuellement. Nous avons aussi suggéré deux heuristiques qui doivent guider ce choix visuel. Dans le futur, plusieurs axes de recherche sont à développer,notamment :

– la déternimation du nombre optimal de coupures ;
– l'aide à la détermination visuelle du nombre de classes distincts le long d'une coupure.

Références

Bock, H. et E. Diday (2000). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag.

Cuvelier, E. et M. Noirhomme-Fraiture (2005). Clayton copula and mixture decomposition. In *ASMDA 2005*, pp. 699–708.

Diday, E. (2002). Mixture decomposition of distributions by copulas. In *Classification, Clustering and Data Analysis*, pp. 297–310.

Diday, E., A. Schroeder, et Y. Ok (1974). The dynamic clusters method in pattern recognition. In *IFIP Congress*, pp. 691–697.

Guo, D. (2003). Coordinating computational and visual approaches for interactive feature selection and multivariate clustering. *Information Visualization 2*(4), 232–246.

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman and Hall.

Nelsen, R. (1999). *An introduction to copulas*. London: Springer.

Poulet, F. (2003). Interactive decision tree construction for interval and taxonomical data. In *Third international workshop on visual data mining - ICDM 2003*, pp. 183–194.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Vrac, M., E. Diday, A. Chédin, et P. Naveau (2001). Mélange de distributions de distributions, décomposition de mélange de copules et application à la climatologie. In *Actes du VIIIème congrés de la Société Francophone de Classification*, pp. 348–355.

Summary

In symbolic data analysis, a complex object can be provided in the form of a continuous distribution. The classification of a set of symbolic objects described by this type of variable, can be obtained by applying a mixture decomposition of archimedean copulas to the values of the distributions calculated in a number Q of distinct points, called cuts. Until now these cuts were arbitrarily selected. In this article we show initially in an empirical way on some examples that the error rate of classification varies with the number of cuts and their positions. We then propose to chose these two parameters thanks to a visual interaction.

Session 4 : Visualisation 2D

Théorie du consensus appliquée au prétraitement des ensembles de données

Edwige Fangseu Badjio, François Poulet

Algorithme interactif pour la sélection de dimensions en détection d'outlier

Lydia Boudjeloud, François Poulet

Tree-View : post-traitement interactif pour des arbres de décision

Nguyen-Khang Pham, Thanh-Nghi Do

Théorie du consensus appliquée au prétraitement des ensembles de données

Edwige Fangseu Badjio, François Poulet

ESIEA Pôle ECD,
Parc Universitaire de Laval-Changé,
38, Rue des Docteurs Calmette et Guérin,
53000 Laval France
{fangseubadjio | poulet}@esiea-ouest.fr

Résumé. Nous présentons une nouvelle approche pour le traitement des ensembles de données de très grande taille en fouille visuelle de données en classification supervisée. Les limites de l'approche visuelle concernant le nombre d'individus et le nombre de dimensions sont connues de tous. Pour pouvoir traiter des ensembles de données de grande taille, une solution possible est d'effectuer un prétraitement de l'ensemble de données avant d'appliquer l'algorithme interactif de fouille visuelle. La réduction du nombre d'individus est effectuée par l'application d'un algorithme de clustering.
La réduction du nombre de dimensions se fait par la combinaison des résultats d'algorithmes de sélection d'attributs par application de la théorie du consensus (avec une affectation visuelle des poids). Nous évaluons les performances de notre nouvelle approche sur des ensembles de données de l'UCI et du Kent Ridge Bio Medical Dataset Repository.

1 Introduction

Nous nous intéressons au problème de prétraitement de grands ensembles de données. Il ressort de l'état de l'art des méthodes de visualisation de données et du diagnostic des systèmes de fouille visuelle de données pour la classification supervisée [Fangseu Badjio et Poulet, 2005] qu'il existe une limite quant à la quantité de données susceptible d'être représentée en une seule fois sur un écran. Pourtant, les progrès scientifiques et techniques permettent aux organisations de stocker des masses de plus en plus importantes de données et d'informations. Il arrive que l'ensemble de données à traiter avec les outils de FVD dépasse la limite tolérée par ces outils, il s'avère alors impossible ou pénible de procéder aux tâches interactives de fouille. En général, les données assez volumineuses comportent des informations bruitées, non significatives, redondantes, etc. Notre but est de réduire les informations contenues dans les ensembles de données volumineux aux informations les plus significatives.

Dans le domaine de l'extraction de connaissances dans les données, il existe des techniques expérimentalement validées pour l'amélioration des résultats des outils d'analyse de données en vue du traitement de grands ensembles de données. Deux approches sont utilisées dans ces techniques : une approche orientée données et une approche orientée algorithme. Nous allons nous intéresser aux approches orientées données pour la sélection d'attributs les plus significatifs de l'ensemble de données à traiter et aux approches orientées

algorithme qui prônent pour la réduction de l'espace de recherche durant cette sélection d'attributs. Un problème majeur se pose alors quant au choix d'une des méthodes connues d'avance pour la sélection d'attributs par exemple, sachant qu'il n'existe pas de méthode qui soit meilleure que toutes les autres dans tous les cas de figure. Une solution qui constitue notre contribution dans ce travail serait d'utiliser une combinaison de techniques ou de stratégies (méthodes de sélection d'attributs). Pour ce faire, nous nous appuyons sur la théorie du consensus dont nous expliciterons le principe dans l'état de l'art dédié à ce sujet. L'utilisation de cette combinaison de stratégies ou d'expertises pour la sélection d'attributs peut être justifiée par l'un des faits suivants :

- il n'est pas possible de déterminer à priori quelle méthode de sélection de sous-ensemble d'attributs est meilleure que toutes les autres (en tenant compte des différences entre le temps d'exécution et la complexité (il s'agit ici de tolérer un temps d'exécution élevé pour un modèle qui nécessite également moins d'attributs)),
- un sous-ensemble optimal d'attributs n'est pas nécessairement unique,
- la décision d'un comité d'experts est généralement meilleure que la décision d'un seul expert.

Comme nous le verrons, l'algorithme de SA proposé qui combine des décisions de plusieurs experts reçoit en entrée des sous-ensembles d'attributs issus de plusieurs expertises et produit comme résultat un sous-ensemble unique d'attributs.

Les résultats obtenus après des expérimentations permettent de conclure que l'approche proposée réduit de façon significative l'ensemble de données à traiter et permet de les traiter interactivement.

Cette contribution commence par un état de l'art et la problématique du sujet abordé, puis, la technique utilisée pour la sélection d'attributs dans des ensembles de données est explicitée, ainsi que des problèmes relatifs à ce traitement. Ensuite, la théorie du consensus, l'algorithme de sélection d'attributs et la méthode d'agrégation des individus contenus dans les ensembles volumineux de données sont présentés. Enfin, nous procédons à des expérimentations avant la conclusion.

2 Etat de l'art et problématique

Il existe plusieurs techniques de visualisation de données multidimensionnelles utilisables en FVD parmi lesquelles on distingue les techniques orientées pixels, les matrices 2 ou 3D, les coordonnées parallèles, etc. Dans la plupart de ces techniques de visualisation (Keim, 1996), le nombre de données susceptibles d'être représentées en même temps à l'écran est limité. Dans un premier temps, nous essayons de résoudre le problème suivant : comment sélectionner des attributs d'un ensemble de données pourvu de plusieurs attributs et rejeter les autres sans nuire à la qualité de l'algorithme utilisé ensuite ? Le sous-ensemble d'attributs qui sera ainsi sélectionné permettra d'obtenir une représentation visuelle beaucoup plus adéquate à la tâche de FVD par rapport à l'ensemble de données en entier.

2.1 Sélection d'attributs significatifs

La sélection d'attributs permet de choisir un sous-ensemble de variables suffisant pour décrire un ensemble de données. C'est un processus permettant d'identifier et de retirer autant que possible les informations redondantes et non utiles de l'ensemble de données. Des

techniques performantes (John et al., 1994), (Kira et Rendell, 1992), etc. de sélection de sous-ensembles d'attributs ont été développées afin de faire face à trois types de problèmes posés par les méthodes d'analyse de données :

- la réduction du coût et la complexité des algorithmes d'apprentissage,
- l'amélioration de la précision des modèles de données obtenus par un processus d'apprentissage,
- l'amélioration de l'intelligibilité de ces modèles de données.

Conformément à l'état de l'art relatif à ce sujet, la sélection d'attributs dans un ensemble de données comprend une phase de génération de sous-ensembles d'attributs, une phase d'évaluation des attributs générés avec une fonction d'évaluation et un critère d'arrêt. La fonction d'évaluation de sous-ensembles d'attributs peut être un algorithme d'induction ou une mesure statistique. Cette fonction permet de distinguer deux types d'approches : des approches de type enveloppe (John et al., 1994) et des approches de type filtre (Kira et Rendell, 1992). Les méthodes existantes de sélection d'attributs peuvent être adaptées pour une utilisation en FVD, mais comme l'indique la section suivante, cette adaptation nécessite de résoudre quelques problèmes au préalable.

2.2 Problèmes en sélection d'attributs significatifs pour la FVD

Comme nous l'avons mentionné antérieurement, il n'existe pas une méthode qui soit meilleure que toutes les autres dans tous les cas. De plus, lorsque le nombre d'attributs de l'ensemble de données à traiter est élevé, la charge cognitive des utilisateurs est grande, sachant qu'il y en aura qui ne pourront même pas réaliser leurs tâches de FVD dans ce contexte. En effet, un environnement de FVD peut être utilisé par : les spécialistes du domaine des données et les spécialistes des méthodes d'analyse de données qui peuvent être soit statisticiens, soit experts en apprentissage automatique.

Il est important d'observer que les différents utilisateurs peuvent être intéressés suivant les cas par les approches filtres et/ou les approches enveloppes. Dans tous les différents cas de figure, un outil d'aide à la sélection d'un sous-ensemble pertinent d'attributs devrait fournir des résultats assez précis. Mais comment retrouver et paramétrer l'algorithme qui suivant le problème à résoudre renverra les meilleurs sous-ensembles d'attributs ? Ceci tout en sachant que :

- la visualisation de plus de deux dizaines d'attributs rend souvent inutilisable la fouille visuelle de données,
- un sous-ensemble optimal d'attributs n'est pas nécessairement unique,
- il n'est pas possible de déterminer à priori quelle méthode de sélection de sous-ensemble d'attributs est meilleure que toute les autres,
- la décision d'un comité d'experts est généralement meilleure que la décision d'un seul expert.

Nous avons défini un nouvel algorithme de sélection d'attributs qui comme nous le verrons combine des décisions pondérées de plusieurs experts (des algorithmes de sélection de sous-ensembles d'attributs). Plus précisément, étant donné deux ou plusieurs méthodes de sélection de sous-ensembles pertinents d'attributs dans un ensemble de données, la question est de savoir comment l'on peut utiliser ces différentes méthodes pour fournir un résultat efficace. Afin de répondre à cette question, nous nous sommes appuyés sur la théorie du consensus qui peut être définie comme un procédé de prise de décision qui utilise entièrement les ressources d'un groupe. Il s'agit d'un champ de recherche qui implique des

procédures dont le but est de combiner plusieurs distributions de probabilités en une seule probabilité dans l'optique de résumer des estimations de plusieurs experts. La théorie du consensus trouve l'une de ses justifications dans le fait qu'une décision prise par un groupe d'experts est meilleure en terme d'erreur quadratique moyenne que la décision d'un seul expert. Une telle démarche possède de nombreux avantages. En effet, statistiquement parlant, la consultation de plusieurs expertises lors de la résolution d'un problème est une façon subjective d'accroître la taille de l'échantillon dans une expérience, un ensemble d'experts permet d'obtenir plus d'information qu'un seul expert (Clemen et Winkler, 1999).

L'algorithme proposé « Consensus Theory Based Feature Selection » (CTBFS) reçoit en entrée des sous-ensembles d'attributs issus de chaque expertise. Une procédure intégrée permet de définir de façon visuelle et interactive des poids à affecter aux décisions de chaque expert. CTBFS retourne en sortie un sous-ensemble d'attributs représentant une agrégation des différents sous-ensembles d'attributs reçus en entrée.

Des représentations graphiques de l'ensemble de données constituées uniquement des attributs sélectionnés sont utilisées pour la définition interactive de poids à affecter aux différents experts qui interviennent dans la sélection d'attributs. Il s'agit ici d'un problème d'optimisation de l'affectation de poids aux experts. Dans un problème d'optimisation, il y a un espace des solutions et une fonction d'évaluation afin d'accéder à la qualité de la solution.

Les sections suivantes présentent la théorie de consensus, l'algorithme de sélection d'attributs basé sur cette théorie ainsi que le processus d'assignation visuelle de poids aux experts.

3 Théorie du consensus : état de l'art

La théorie du consensus consiste à rechercher un accord parmi des solutions proposées par un groupe d'experts. Cette théorie a été largement utilisée en classification, statistiques et en sciences sociales (Barthélemy et al., 1984), (Day et McMorris, 2003), etc. Selon (Clemen et Winkler, 1999), la consultation de plusieurs experts constitue une version subjective d'augmentation de la taille de l'échantillon dans une expérience. Ces experts peuvent en effet fournir plus d'information qu'un seul expert.

Les méthodes basées sur le consensus (combinaison ou agrégation) peuvent être classées en deux catégories : les approches mathématiques et les approches comportementales.
Les approches comportementales tentent de générer un agrément entre les experts par une interaction entre eux (Clemen et Winkler, 1999).

Dans les approches mathématiques (Chen et al., 2005), les opinions individuelles d'experts sont exprimées sous forme de distributions de probabilité subjectives d'un événement incertain et sont combinés par diverses méthodes mathématiques pour former une distribution de probabilité agrégée. Il existe plusieurs modèles de combinaison mathématiques pour la définition d'un consensus (Winkler, 1968), (French, 1985), etc. Les approches utilisées dans ces combinaisons peuvent être axiomatiques ou bayésiennes. Les approches utilisées de façon usuelles sont basées sur des axiomes, comme le « linear opinion pool (Lin-OP)» ou le « logarithmic opinion pool (Log-OP)».

Le Lin-OP est la somme linéaire des probabilités à posteriori de chaque solution experte. La fonction de décision utilisée à cet effet est la suivante :

$$p(\theta) = \sum_{i=1}^{n} w_i \, p_i(\theta) \quad (1)$$

où $p_i(\theta)$ est la distribution de probabilité d'un événement incertain θ, $p(\theta)$ représente la distribution de probabilité agrégée, et w_i le facteur poids, avec $\sum_{i=1}^{n} w_i = 1$.

Le Log-OP est la moyenne pondérée géométrique des distributions de probabilités individuelles. La fonction de décision dans ce cas est :

$$p(\theta) = k \prod_{i=1}^{n} p_i(\theta)^{w_i} \quad (2)$$

où k est une constante de normalisation permettant de s'assurer que l'opinion agrégée est une distribution de probabilité.

Dans la fonction de décision présentée ci-dessus, le facteur poids détermine l'influence de chaque expert sur la décision commune. Il existe deux types d'opérateurs d'affectation de poids : les opérateurs contextuels et les opérateurs non contextuels. Nous proposons l'utilisation d'une fonction dépendante du contexte de la décision à prendre pour l'affectation de poids aux différentes expertises. Cette méthode basée sur des représentations graphiques utilise comme nous le verrons les capacités usuelles humaines en perception.

4 Algorithme de sélection d'attributs basé sur la théorie du consensus (CTBFS)

Le domaine considéré est constitué d'une valeur limite du nombre d'attributs susceptibles d'être correctement visualisés et traités da façon interactive (C_{cmd}), un ensemble M d'experts (algorithmes de sélection d'attributs) $E = \{E_1, \ldots, E_M\}$, chaque expert E_i dispose d'un sous-ensemble de L experts (qui représentent les différents critères ou paramètres importants des algorithmes de sélection d'attributs) $E_i = \{e_1, \ldots, e_l\}$. L'utilisation de ce sous-ensemble d'experts (E_i) peut être justifié par le fait que dans un algorithme de sélection d'attributs significatifs, il n'y a aucun critère qui permet d'obtenir de meilleurs résultats que tous les autres. Chaque critère possède des attributs de qualité spécifiques. Il est nécessaire de prendre en considération tous les différents attributs de qualité.

Nous avons aussi un sous-ensemble d'attributs $DS = \{D_1, \ldots, D_L\}$, où $D_i = \{d_1, \ldots, d_K\}$ et K est variable. Les sous-ensembles d'attributs sont disponibles selon les paires expert/attributs (e_j, D_j), où $e_j \in E_i$ et $D_j \in DS$.

Chaque attribut sélectionné par un sous expert e_j a une fréquence $freq = 1/nb$ d'apparition dans la décision finale, où nb est le nombre d'attributs sélectionnés par le sous expert.

Nous définissons un critère de préférence d'un attribut (règle de consensus) comme étant le produit des fréquences d'apparition de l'attribut dans les sous-ensembles d'attributs des experts. Nous utilisons la Log-OP pour le calcul de la préférence d'un attribut d.

$$pref(X = d) = \prod_{i=1}^{N} P(X = d \mid D_i = b_i)^{w_i} \quad (3)$$

où : $P(X = d \mid D_i = b_i)$ est la probabilité à posteriori que l'attribut testé appartienne

au sous-ensemble d'attributs à sélectionner lorsque la décision du $m^{ième}$ expert est b_i, w_i est le poids assigné à l'expert.

A cette étape, il est important de revenir sur l'affectation de poids aux différents experts intervenant dans la procédure de sélection des sous-ensembles d'attributs.

4.1 Affectation visuelle de poids pour la prise de décision collective

Une décision collective constitue une décision raisonnable que tous les membres d'un groupe peuvent accepter. Les fonctions de combinaison ou d'agrégation d'opinion nécessitent un facteur poids (voir les formules 1 et 2). Le poids détermine l'influence de chaque expert sur la décision commune. Les poids affectés aux différents experts peuvent être égaux ou on peut rechercher une combinaison optimale de facteurs poids. Toute la difficulté relative à un tel processus est de trouver la stratégie d'optimisation de ces facteurs, sachant que notre objectif est de retrouver des poids qui permettent de réduire de façon significative le nombre d'attributs et si possible d'avoir une meilleure précision en fouille visuelle de données (FVD).

Les poids affectés aux experts doivent à cet effet être proportionnels à leurs décisions. L'idée ici est de donner des poids élevés aux meilleurs experts (en terme de représentation graphique de leur sélection d'attributs).

Nous pensons que la meilleure façon de juger de la qualité de l'expertise proposée par les différentes méthodes de sélection d'attributs serait de procéder à des représentations graphiques de l'ensemble de données à traiter avec uniquement les attributs les plus significatifs choisis par chaque expert. L'idée tout au long de ce processus rappelons-le est de donner un poids faible ou alors de ne pas tenir compte de la décision d'un expert qui aurait choisi un très grand nombre d'attributs (d'où une impossibilité de représenter graphiquement l'ensemble de données).

La méthode d'affectation de poids que nous proposons a pour fondements théoriques un principe de la théorie de Gestalt (une vue d'ensemble est meilleure que la somme des parties) et des propriétés pré-attentives de la vision humaine. En ce qui concerne le principe de Gestalt, en visualisant l'ensemble d'éléments intervenant dans une décision, un processus cognitif se met en place.

Dans notre contexte, l'application du principe de Gestalt en ce qui concerne la visualisation de l'ensemble d'éléments rentrant dans le processus de décision se résume en une représentation graphique multi vue. Chaque vue représente le point de vue de chaque expert, c'est-à-dire la représentation graphique de l'ensemble de données pourvu uniquement des attributs sélectionnés par l'expert, comme l'indique la figure 1.

En effet, la technique utilisée pour l'affectation visuelle de poids aux experts intervenant dans le processus de décision collective est une représentation graphique à vues multiples des coordonnées parallèles (Inselberg, 1985). Chaque vue représente l'ensemble de données à traiter réduit par un des experts. Les coordonnées parallèles permettent de représenter en 2D des données multidimensionnelles sans perte d'information.

Six experts de type filtre ont servi à la sélection des attributs visualisés dans la figure 1. L'expert 1 représente le critère de sélection consistance, l'expert 2 représente l'entropie de Shannon, l'expert 3 quant à lui utilise la distance comme fonction d'évaluation. La fonction d'évaluation pour l'expert 4 est le gain d'information, le coefficient de Gini pour l'expert 5 et le coefficient de Cramer pour l'expert 6.

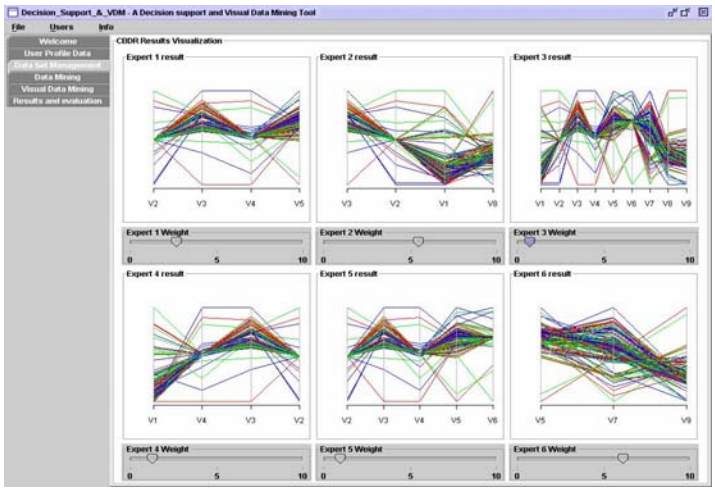


FIG. 1 – *Outil d'affectation visuelle de poids aux experts intervenant dans CTBFS.*

Il est à noter que les outils usuels d'affectation de poids sont des « boîtes noires ». L'avantage principal de l'approche ainsi proposée tient du fait que l'utilisateur est impliqué et participe dans le processus de prise de décision. Il existe un ensemble de propriétés visuelles qui sont traitées de manière pré attentive très rapidement, avec précision et sans effort particulier. Ce qui permet aux utilisateurs d'affecter des poids convenables aux différents experts.

De plus, les techniques de visualisation permettent d'améliorer la résolution de problèmes. La visualisation permet de découvrir plus aisément des motifs dans les données, de réduire l'espace de recherche d'information par rapport aux méthodes automatiques, de procéder à des opérations perceptuelles d'inférence et d'augmenter la mémoire et les ressources de traitement de l'utilisateur (Dull et Tegarden, 1999), (Card et al., 1999) et (Tegarden, 1999).

4.2 Réduction du nombre d'observations

Certains ensembles de données disposent d'un grand nombre d'attributs et/ou d'observations. Nos investigations en ce qui concerne la réduction des observations dans un ensemble de données consistent à agréger l'information contenue dans cet ensemble de données.

En effet, au lieu de traiter l'ensemble de données avec un grand nombre d'individus, l'idée est d'utiliser juste un échantillon S des individus de cet ensemble de données.
Considérons une collection d'observations $\{D_1, \ldots, D_n\}$, à représenter graphiquement, qui nécessite aussi l'application des procédures de FVD. L'agrégation S de cette collection d'observations est une partition $\{S_1, \ldots, S_k\}$ de $\{D_1, \ldots, D_n\}$ et tout S_i est un cluster. Le clustering divise les observations d'un ensemble de données en groupes pour des besoins d'agrégation ou pour une amélioration de la compréhension de ces données. Le clustering qui a été utilisé en compression de données permet de retrouver très efficacement les plus proches voisins d'un point. Pour le clustering, l'ensemble de données initial est séparé en

observations de même classe class (ID_i). Ensuite, pour chaque ensemble ID_i, nous appliquons l'algorithme K-means (MacQueen, 1967) afin de retrouver les clusters ou groupes d'éléments disponibles dans ID_i. Typiquement, un algorithme de clustering permet de partitionner N entrées $x_1, x_2, \ , x_N$ en k clusters. Les objets regroupés dans chaque groupe résultant sont similaires entre eux et différents des objets des autres groupes. Les algorithmes de clustering essayent de trouver une partition k qui maximise une fonction d'objectif en ce qui concerne la mesure de similarité. Par exemple, une fonction d'objectif peut trouver le cluster qui maximise la somme des similarités des objets de la même partition (cluster).

Cette approche a déjà été utilisée dans le cadre d'un pré-traitement de grands ensembles de données avec des algorithmes de type « support vector machine (SVM)» par (Do et Poulet, 2005), (Poulet, 2005). Les auteurs ont testé et validé cette approche sur de grands ensembles de données.

5 Expérimentations

Pour les besoins d'expérimentation de la technique proposée qui a été développée sous Windows avec Java et le langage R, nous utilisons un pentium IV, 1.7 GHz. Les ensembles de données que nous utilisons proviennent de l'UCI (Blake et Merz, 1998) et du Kent Ridge Bio-medical Data Set Repository (Jinyan et Huiqing, 2002). Cette étude de cas dispose de deux tests. Pour les besoins de ces expérimentations, les poids affectés aux différents experts ont pour valeur 1.

5.1 Sélection d'attributs

5.1.1 Premier test

Le domaine considéré dans le cadre de cette première expérimentation est constitué d'un ensemble M constitué de 3 experts de type filtre et de 3 experts de type enveloppe E = *{consistence, entropie de Shannon, distance, (LDA, QDA, Kppv)* (Ripley, 1996)*}*, le nombre d'attributs susceptibles d'être traités convenablement est $C_{cmd} = 20$.

Les résultats de l'algorithme proposé (CTBFS) sont comparés à ceux de Las Vegas Filter (Liu et Setiono, 1996), un algorithme de sélection d'attributs de type filtre et StepClass du package KlaR (langage de programmation R), un algorithme de sélection d'attributs de type enveloppe. A cet effet, nous évaluons les performances des ensembles de données pourvus des attributs sélectionnés par ces trois méthodes (LVF, StepClass et CTBFS) avec l'algorithme des k plus proches voisins kppv (implémentation de WEKA (Witten et Eibe, 2005)). Nous avons fixé le paramètre K de l'algorithme des kppv à 1.

Les ensembles de données à traiter dans le cadre de cette première expérimentation sont pourvus de nombreux attributs (colonne 2 de TAB.1) et qu'il serait impossible de les visualiser en une seule fois à l'écran quelque soit la méthode de représentation graphique choisie.

Les résultats exposés dans TAB. 1 permettent d'observer que l'algorithme CTBFS que nous proposons permet de réduire considérablement le nombre d'attributs des ensembles de données comme le montre les résultats de la colonne 3 de TAB. 1. La colonne 5 de ce tableau quant à elle fait observer que la précision de l'algorithme de kppv est améliorée pour 4 ensembles de données sur 7. Pour les trois autres ensembles de données, on assiste certes à

une perte de précision avec un écart maximal de 16.97% avec un minimum de précision de 68.87% mais l'ensemble de données final peut être visualisé et traité de manière interactive. Ce qui n'est pas le cas des ensembles de données initiaux comme nous l'avons souligné.

Nom	NbAt_Initial	NbAt_CTBFS	Précision_initiale	Précision_CTBFS
Lung-Cancer	57	**4**	37.5%	**75%**
Promoter	59	**9**	**85.84%**	68.87%
Sonar	60	**8**	**86.54%**	71.15%
Arrhythmia	280	**4**	53.44%	**59.96%**
Isolet	618	**14**	**85.57%**	70.24%
ColonTumor	2000	**19**	77.42%	**79.03%**
CentralNervSyst	7129	**20**	56.67%	**60%**

TAB. 1 – *Comparaison du nombre d'attributs et de la précision obtenus avec l'algorithme des kppv avant et après la sélection d'attributs par l'algorithme CTBFS*

Nom	NbAttr CTBFS	NbAttr LVF	NbAttr Stepclass	CTBFS précision	LVF précision	Stepclass précision
Lung-Cancer	**4**	17	**4**	**75%**	62.5%	71.87%
Promoter	**9**	16	59	68.87%	80.19%	**85.85%**
Sonar	**8**	18	**4**	***71.15%***	82.21%	***71.63%***
Arrhythmia	**4**	109	**4**	**59.96%**	54.65%	**60.84%**
Isolet	**14**	268	**8**	***70.24%***	83%	***57.98%***
ColonTumor	**19**	918	**5**	**79.03%**	77.42%	**79.03%**
CentralNervSyst	**20**	3431	**8**	***60%***	***58.33%***	**71.67%**

TAB. 2 – *Comparaison du nombre d'attributs et de la précision obtenus avec l'algorithme des kppv avant et après la sélection d'attributs par les algorithmes CTBFS, LVF et Stepclass.*

On observe à travers la colonne 3 de TAB. 2 que la méthode LVF permet de sélectionner un nombre très important d'attributs, qu'il serait impossible de visualiser (par exemple pour les ensembles de données Arrhythmia, Isolet, ColonTumor et CentralNervSyst). Par rapport à la méthode proposée, la précision obtenue pour ces ensembles de données est équivalente voire supérieure par exemple pour l'ensemble de données Isolet, sachant que l'algorithme CTBFS renvoie au maximum 20 attributs. En ce qui concerne l'algorithme Stepclass, l'ensemble de données Promoter possède aussi un nombre important d'attributs.

En terme de précision, en dehors de l'ensemble de données Promoter pour lequel CTBFS a une précision inférieure à celle de Stepclass et de LVF, la précision obtenue pour les autres ensembles de données avec l'algorithme proposé est au moins égale suivant les cas à celle de LVF ou à celle de Stepclass mais avec un nombre d'attributs qui convient à la fouille visuelle de données.

5.1.2 Deuxième test

Dans ce second test, on va s'intéresser au comportement de l'algorithme CTBFS sur des ensembles de données de taille moyenne. Le domaine considéré reste le même. Les résultats obtenus par CTBFS sont aussi comparés à ceux de LVF et Stepclass. Dans cette expérimentation, nous évaluons les performances des ensembles de données pourvus des

attributs sélectionnés par LVF, StepClass et CTBFS avec l'algorithme C4.5, implémentation de WEKA.

	NbAt Initial	NbAt CTBFS	NbAt LVF	NbAt STEP	Précis CTBFS	Précis LVF	Précis STEP
arrhythmia	280	4	**109**	4	**66.15%**	**66.15%**	63.72%
bupa	6	5	2	4	**68.99%**	52.17%	60%
credit_a	15	5	3	4	**86.53%**	73.06%%	74.90%
crx	15	5	3	5	**77.97%**	63.33%	73.48%
glass	9	3	2	2	**61.22%**	47.66%	**62.62%**
hepatitis	19	6	4	16	**80%**	79.35%	**80.65%**
ionosphere	34	4	8	2	**88.89%**	83.76%	79.77%
isolet	618	14	**268**	8	**66.58%**	**73.83%**	58.63%
lung_cancer	57	4	17	4	**71.88%**	62.5%	65.63%
monks	6	4	3	2	**89.52%**	74.19%	72.58%
promoter	59	9	16	**59**	74.53%	68.87%	**79.25%**
sonar	60	8	18	4	**71.15%**	64.90%	65.87%
Voting	16	7	3	8	**94.94%**	88.51%	**96.32%**

TAB. 3 – *Comparaison du nombre d'attributs et de la précision obtenus avec l'algorithme C4.5 avant et après la sélection d'attributs par les algorithmes CTBFS, LVF et Stepclass.*

Le premier objectif du prétraitement des données pour la FVD rappelons-le est la réduction du nombre d'attributs, autrement, il est impossible de traiter l'ensemble de données. Ensuite, on s'intéresse à la variation de la précision dans les ensembles de données résultant de ce prétraitement. TAB. 3 à la colonne 3 montre que ce premier objectif est atteint pour les différents ensembles de données testés. Il est à noter que l'observation des résultats obtenus avec l'algorithme LVF relève un nombre beaucoup plus important d'attributs sélectionnés pour les ensembles de données Arrythmia et Isolet.

En ce qui concerne la précision, on observe un gain avec l'approche proposée sur plusieurs ensembles de données traités (bupa, credit_a, crx, ionosphere, lung_cancer, monks et sonar). Une égalité de précision apparaît entre CTBFS et LVF/Stepclass pour les ensembles de données arrythmia/hepatitis. Etant donné le nombre d'attributs sélectionnés par LVF pour Isolet et Stepclass pour Promoter, nous pouvons conclure que CTBFS permet d'obtenir de meilleurs résultats, le traitement interactif pouvant s'opérer dans les deux cas de figure avec cette méthode.

6 Conclusion

Nous avons présenté un algorithme basé sur la théorie du consensus et l'affectation visuelle de poids pour la sélection d'attributs significatifs en FVD. En effet, lorsque le nombre d'attributs et/ou le nombre d'observations d'un ensemble de données est important, il s'avère impossible ou alors pénible de représenter graphiquement l'ensemble de données et d'observer des corrélations dans cet ensemble de données.

La technique présentée permet de définir un nombre maximum d'attributs à sélectionner dans l'ensemble de données à traiter, nombre rendant possible la visualisation de ces données. La première nécessité pour nous est de pouvoir représenter visuellement l'ensemble

de données à traiter. Les expérimentations effectuées à cet effet ont été concluantes. Ensuite, nous nous sommes intéressés à la précision des algorithmes C4.5 et kppv sur les ensembles de données à traiter pourvus uniquement des attributs relevés par application de la théorie du consensus. Force a été pour nous de constater que pour plusieurs de ces ensembles de données le taux de précision était amélioré par rapport au taux de précision initial (pour les kppv) et par rapport à LVF et Stepclass pour C4.5. Cette comparaison a été concluante comme l'indique les résultats obtenus en section 5. A la suite de la sélection des attributs, l'utilisation des algorithmes de clustering nous permet de réduire le nombre d'individus des ensembles de données de 50 à 75% avec un maximum de 200 clusters par application de l'algorithme K-Means. Faute d'espace, les expérimentations de ce processus n'ont pas été présentées dans cette contribution.

Comme perspectives à ces travaux, nous comptons étendre l'application de la théorie du consensus au choix de la meilleure méthode de classification supervisée ou non supervisée pour un ensemble de données à traiter.

Références

J.-P. Barthélemy, B. Leclerc et B. Monjardet (1984). Quelques aspects du consensus en classification, in Data analysis and informatics (eds. Diday et al.), Amsterdam: Elsevier, 307-315.

C. Blake and C. Merz (1998). UCI Repository of machine learning databases. Irvine, University of California, Department of Information and Computer Science, from www.ics.uci.edu/~ mlearn/MLRepository.html.

S.K. Card, J. D. Mackinlay, and B. Shneiderman (1999). Information Visualization: Using Vision to Think. Academic Press.

Y. Chen, C.-H. Chu, T. Mullen, and D.M. Pennock (2005). Information Markets vs. Opinion Pools: An Empirical Comparison , ACM Conference on Electronic Commerce (EC 05), Vancouver, British Columbia, Canada, June 5-8, 2005:58-67.

R. T. Clemen and R.L. Winkler (1999). Combining probability distributions from experts in risk analysis. Risk Analysis, 19(2):187–203, 1999.

W.H.E. Day and F.R. McMorris (2003). Axiomatic Consensus Theory in Group Choice and Biomathematics, SIAM, Philadelphia.

T-N. Do and F. Poulet (2005). Mining Very Large Datasets with SVM and Visualization, in proc. of ICEIS'05, 7th Int. Conf. on Entreprise Information Systems, Miami, USA, 2005, Vol. 2, pp.127-141, 2005.

R.B. Dull, D.P.A Tegarden (1999). Comparison of three visual representations of complex multidimensional accounting information. Journal of Information Systems. Vol. 13, No. 2 (Fall), pages 117-131, 1999.

C.F.Eick, M. Zeidat Nidal, Zhao Zhenghong (2004). Supervised Clustering - Algorithms and Benefits. ICTAI 2004, pages 774-776.

E. Fangseu Badjio, F. Poulet Towards usable visual data mining environments, In proc. of HCII'05, the 11th International Conference on Human-Computer Interaction, Las Vegas, Nevada, USA, Jul 2005.

S. French (1985). Group consensus probability distributions: a critical survey. Bayesian Statistics, 2:183–202, 1985.

A. Inselberg (1985). The plane with parallel coordinates. The Visual Computer, 1, pages 69--91.

L. Jinyan and L. Huiqing (2002). Kent Ridge Bio-medical Data Set Repository. http://sdmc.lit.org.sg/GEDatasets.

G. H. John, R. Kohavi and K. Pfleger (1994). Irrelevant features and the subset selection problem, in `International Conference on Machine Learning', pp. 121-129.

D.A. Keim (1996). Pixel-oriented Visualization Techniques for Exploring Very Large Databases. Journal of Computational and Graphical Statistics, March 1996.

K. Kira and L. A. Rendell (1992). A practical approach to feature selection. In Proc. of the Tenth Int'l Conf. on Machine Learning, pages 500–512.

H. Liu and R. Setiono (1996). A probabilistic approach to feature selection: a filter solution. In Proc, The 13th International Conference on Machine Learning, pages 319-327.

J. MacQueen (1967) Some methods for classification and analysis of multivariate observations. In Le Cam, L. M. and Neyman, J., editors, Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, pages 281-297, Berkeley, Califonia. University of California Press.

F. Poulet (2005). Visual SVM, in proc. of ICEIS'05, 7th Int. Conf. on Entreprise Information Systems, Miami, USA, 2005, Vol. 2, pp.309-314, 2005.

Quinlan J. R. *C4.5: Programs for Machine Learning*. Morgan-Kaufman, San Mateo, CA, 1993.

B. D. Ripley (1996). Pattern Recognition and Neural Networks. Cambridge University Press.

D.P. Tegarden (1999). Business information visualization. Communications of the AIS. Vol. 1, Article 4 (January).

R.L. Winkler (1968). The consensus of subjective probability distributions. Management Science, 15(2):B61–B75.

I.H.Witten and F. Eibe (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

Summary

Visualization methods do not scale well with high number of features and/or observations. We present an approach using a consensus theory based feature selection (CTBFS) algorithm, clustering for sampling and visualization for weight assignment in order to aggregate multivariate and multidimensional datasets.

Algorithme interactif pour la sélection de dimensions en détection d'outlier

Lydia Boudjeloud, François Poulet
ESIEA Pôle ECD
38, rue des docteurs Calmette et Guérin
Parc Universitaire de Laval-Changé
53000 Laval
boudjeloud|poulet@esiea-ouest.fr

Résumé. Nous présentons un algorithme génétique semi-interactif de sélection de dimensions dans les grands ensembles de données pour la détection d'individus atypiques (outliers). Les ensembles de données possédant un nombre élevé de dimensions posent de nombreux problèmes aux algorithmes de fouille de données, une solution souvent retenue pour le traitement de tels ensembles de données est d'effectuer un pré-traitement afin de ne retenir que les dimensions "intéressantes". Mais le nombre de solutions potentielles est trop important pour pouvoir les passer toutes en revue. Une solution est alors d'utiliser un algorithme génétique pour le choix du sous-ensemble de dimensions à retenir. Par ailleurs nous souhaitons donner un rôle plus important à l'utilisateur dans le processus de fouille, pour cela nous avons choisi d'utiliser un algorithme génétique interactif. Dans ce type d'approche c'est l'utilisateur qui remplace la fonction d'évaluation de l'algorithme génétique. Notre approche n'élimine pas complètement cette fonction d'évaluation mais la couple avec l'évaluation de l'utilisateur, on parle alors d'algorithme génétique semi-interactif. Enfin, l'importante réduction du nombre de dimensions nous permet de visualiser les résultats de l'algorithme de détection d'outlier. Cette visualisation permet à l'expert des données d'étiqueter les éléments atypiques, par exemple est-ce que ce sont des erreurs ou simplement des individus différents de la masse.

1 Introduction

Nous nous intéressons à la recherche d'outliers (individus atypiques) dans les ensembles de données ayant un grand nombre de dimensions. Pour pouvoir traiter de tels ensembles de données (par exemple les ensembles de données de fouille de texte ou de bio-informatique), la plupart des algorithmes de fouille de données actuels nécessitent un prétraitement permettant de réduire le nombre de dimensions (avec plus ou moins de perte d'information). L'approche la plus intuitive pour appréhender le problème des grandes dimensions est d'énumérer tous les sous-ensembles de dimensions possibles et de rechercher le sous-ensemble qui satisfait la problématique traitée. Cependant, le fait d'énumérer (rechercher) toutes les combinaisons possibles est un problème NP-difficile (Narenda et Fukunaga, 1977). Parmi les solutions proposées pour ce problème, on retrouve la réduction de dimensions (combinaison de dimensions, généralement linéaire) et la sélection de dimensions (on n'utilise qu'un sous-ensemble des dimensions originales). L'avantage de cette dernière solution est que nous ne perdons pas l'information que pourrait apporter la dimension, car

elle est considérée individuellement non en combinaison (linéaire) avec d'autres dimensions. Les techniques de sélection de dimensions consistent donc à réduire l'ensemble des dimensions considérées. L'objectif est de réduire la complexité, augmenter la précision de la prédiction et/ou réduire le temps de traitement des données, en sélectionnant le sous-ensemble de dimensions de taille minimale (Dash et al., 1997). L'étude du problème de sélection de dimensions se justifie facilement par le fait qu'une recherche exacte a un coût exponentiel en temps de calcul et en espace mémoire. En effet, la sélection d'un sous-ensemble de dimensions demanderait l'exploration de tout l'espace de recherche. Pour $|D|$ dimensions, la recherche exhaustive consiste à explorer $2^{|D|}$-1 sous-ensembles possibles. La recherche d'un sous-ensemble de s dimensions parmi D consiste à appliquer le critère d'évaluation $C^s_{|D|}$ fois. Si on trouve S' ensembles, on aura donc une complexité de $\sum_{s=0}^{S'} C^s_{|D|} = O(|D|^s)$. Lorsque l'on s'attaque à des problèmes réels, il faut se résoudre à un compromis entre la qualité des solutions obtenues et le temps de calcul utilisé. Au milieu des années 1970 sont apparues des méthodes qui supervisent l'évolution de solutions fournies par des heuristiques. Ces méthodes assurent un compromis entre diversification (quand il est possible de déterminer que la recherche se concentre sur de mauvaises zones de l'espace de recherche) et intensification (on recherche les meilleures solutions dans la région de l'espace de recherche en cours d'analyse). Ces algorithmes ont été appelés métaheuristiques et ont pour objectif de trouver des solutions dont la qualité est au-delà de ce qu'il aurait été possible de réaliser avec une simple heuristique (Jourdan, 2003). Dans cet article, nous proposons un algorithme génétique pour le choix du sous-ensemble de dimensions à retenir.

Par ailleurs nous souhaitons donner un rôle plus important à l'utilisateur dans le processus de recherche et de sélection de l'algorithme génétique, pour cela nous avons choisi d'utiliser un algorithme génétique interactif. Nous présentons donc une nouvelle méthode interactive, proposant elle-même des solutions potentielles à l'utilisateur. Les solutions de l'algorithme génétique se présentent sous forme de sous-ensembles de dimensions. Puisque le nombre de dimensions utilisé est faible, on peut ensuite visualiser les éléments de l'ensemble de données sur ces sous-espaces de dimensions (à l'aide de matrices de scatter-plot (Carr et al., 1987) ou de coordonnées parallèles (Inselberg, 1985)) pour permettre à l'expert d'interpréter les résultats obtenus. Nous présentons donc des visualisations d'un ensemble de données projeté sur quelques sous-ensembles de dimensions à l'utilisateur, ce dernier pourra lui même juger de la pertinence de la visualisation présentée selon ses objectifs (repérer les dimensions les plus pertinentes pour la détection d'outlier). Pour cela, nous avons choisi d'utiliser un algorithme génétique interactif (AGI). D'une manière générale, cet algorithme fonctionne de la façon suivante : une première évaluation automatique se fait à l'aide d'un critère d'évaluation des sous-espaces de dimensions pour la détection d'outlier basé sur les distances, les données sont ensuite visualisées et présentées graphiquement à l'utilisateur pour une seconde évaluation visuelle. Ce dernier choisit celles qui lui semblent les plus pertinentes. Les caractéristiques visuelles des sous-espaces de données sélectionnés sont prises en compte par l'algorithme pour la génération suivante de sous-espaces, qui sont à nouveau présentés à l'utilisateur et ainsi de suite jusqu'à ce que la recombinaison des caractéristiques permette de générer une projection visuelle de données complètement satisfaisante pour l'utilisateur. Un des avantages de cette méthode est de faire collaborer deux méthodes, automatique et visuelle. La première automatique, à l'aide des critères

d'évaluation permet d'éliminer les solutions redondantes ou bruitées et la seconde visuelle et interactive permet à l'utilisateur de participer au processus de recherche et d'aborder un aspect d'évaluation des solutions présentées sous forme visuelle qui représente justement un nouveau domaine de recherche. Un autre avantage est que la méthode s'adresse plus particulièrement au spécialiste des données qui peut utiliser les connaissances du domaine pour l'interprétation visuelle des résultats tout au long du processus de fouille et ainsi apporter un aspect d'aide à la décision. La méthode lui permet d'étiqueter les éléments atypiques, par exemple est-ce que ce sont des erreurs ou simplement des individus différents de la masse. Nous détaillons dans une première partie notre algorithme puis commentons les résultats obtenus sur quelques ensembles de données, nous essayerons ensuite d'interpréter visuellement les résultats obtenus, puis nous terminons par la conclusion et les travaux futurs.

2 Algorithme : Viz-IGA

Face au problème de la prise en compte des préférences de l'utilisateur, des auteurs ont montré comment ce dernier pouvait sélectionner lui même directement les solutions qui le satisfont le plus, sans passer par une fonction automatique parfois impossible à définir (Takagi, 2001), (Hayashida et Takagi, 2000). Une nouvelle catégorie d'algorithmes génétiques est née, connue sous le nom d'Algorithmes Génétiques Interactifs. Ces algorithmes génétiques interactifs (AGIs) permettent ainsi des applications nouvelles comme l'obtention de belles images de synthèse ou de sons polyphoniques. Dans ces applications l'utilisateur note selon ses critères les individus qui représentent des images (ou des sons) et l'AGI fait évoluer les individus selon les préférences de l'utilisateur. Les algorithmes génétiques interactifs (AGIs) sont une extension des AGs dans lesquels a lieu une interaction entre la méthode de recherche et l'utilisateur, ce dernier guidant la méthode vers les solutions ayant les caractéristiques qu'il préfère. L'algorithme génétique standard doit être modifié comme suit : les étapes d'évaluation et de sélection automatique des individus sont remplacées par une présentation des individus à l'utilisateur qui sélectionne en un certain nombre. Cela implique notamment de limiter la population à un petit nombre d'individus. Les principales conditions d'application des AGIs citées précédemment sont particulièrement intéressantes dans notre cas d'évaluation visuelle des sous-espaces de dimensions sélectionnés. En effet, l'interprétation visuelle des résultats obtenus est importante pour la détection d'outlier et valider des sous-espaces qui présentent mieux les solutions attendues est aussi une étape importante dans le processus de recherche de sous-ensembles de dimensions. L'utilisateur peut ainsi intervenir dans le processus de recherche des sous-espaces de dimensions pertinents.

2.1 Initialisation

Nous considérons que l'utilisateur veut avoir des représentations graphiques de données dans un sous-espace de dimensions sur lesquels il peut voir des outliers facilement détectables. Le but recherché est d'aider l'utilisateur à comprendre et interpréter ses données à travers les résultats de l'algorithme. Pour cela, on va lui proposer des représentations graphiques en *k-D* des données avec une des méthodes de visualisation de données

(coordonnées parallèles (Inselberg, 1985), matrices de scatter-plot (Carr et al., 1987) ou star plot (Card et al., 1999)).

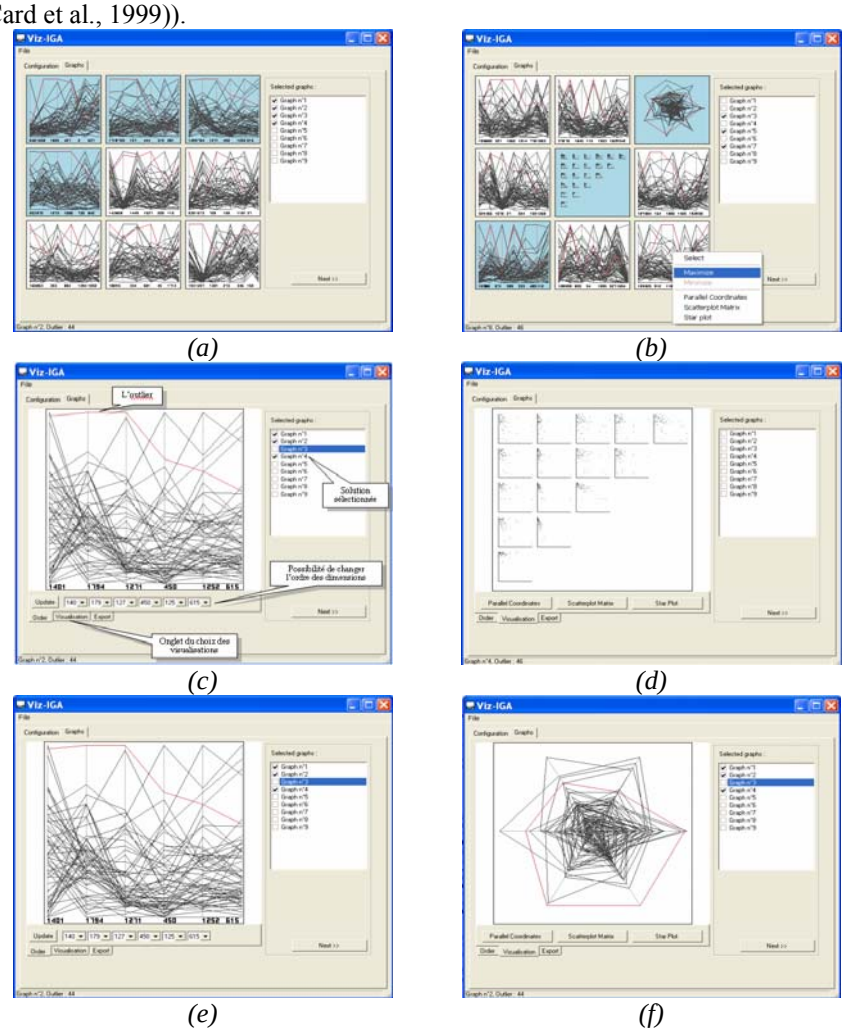


FIG. 1 – *Différentes interfaces de Viz-IGA.*

Le choix de *k*, de l'ensemble de données à traiter et de la méthode de visualisation revient à l'utilisateur. Nous proposons ces méthodes de visualisation car elles permettent à l'utilisateur d'interagir avec l'ensemble des données sous la forme d'une série de projections, l'utilisateur peut voir la pertinence d'une dimension et le comportement des éléments de l'ensemble de données.

2.2 Représentation de l'individu et opérateurs génétiques

Un individu est donc un sous-ensemble de dimensions représenté par une combinaison d'axes (Axe_1, Axe_2, …, Axe_k), ce sous-ensemble étant sélectionné par un algorithme

génétique (Boudjeloud et Poulet, 2005). Le codage choisi consiste à fixer un nombre *s* (taille du sous-ensemble de dimensions à sélectionner), ainsi un individu de l'AG (ou de l'AGI) représente une combinaison possible de *s* dimensions. Ce type de codage permet une meilleure interactivité avec les dimensions. La taille *s* est un paramètre d'entrée de l'algorithme génétique. Pour notre problème nous nous basons sur des petites tailles pour faciliter l'interprétation visuelle des résultats (typiquement inférieur à 10). L'opérateur de croisement échange des sous-groupes de dimensions des individus parents, en respectant la contrainte de non présence de clones (les individus qui ont le même sous-ensemble de dimensions, présentées dans un ordre différent ou pas, sont interdits dans notre algorithme de même que des dimensions identiques dans un même individu). L'opérateur de mutation échange aléatoirement un gène en respectant les mêmes conditions. Nous avons opté pour un point de coupe "optimisé" (Boudjeloud et Poulet, 2005), où l'on détermine dans ce cas le meilleur point avant d'opérer la coupe, ce qui implique une évaluation de chaque individu issu de chaque coupe possible.

2.3 Evaluation visuelle semi-interactive

Pour que l'utilisateur puisse évaluer visuellement les individus de la population et avoir des représentations visuelles de sous-espaces de données sur lesquelles il peut voir des outliers facilement détectables, nous présentons les individus à l'écran (figure 1-a) en faisant une première évaluation automatique à l'aide d'un critère d'évaluation à base de distance (Boudjeloud et Poulet, 2005). Neuf individus choisis aléatoirement dans la population de l'AG sont affichés simultanément pour ne pas surcharger l'interface et faciliter les comparaisons. Pour évaluer la qualité d'un individu, l'expert dispose de la représentation en coordonnées parallèles qu'il peut éventuellement changer en matrices de scatter-plot ou star-plot, selon ses préférences (figure 1-b). Il ne faut pas oublier que nous traitons des ensembles de données de grandes dimensions (de l'ordre de dix à cent milles dimensions, cf. figure 4), notre objectif principal est d'obtenir des visualisations de données pas trop surchargées, pour cela l'AG traite et présente des visualisations de données avec des sous-ensembles de petite taille (de 4 à 10 pour que les visualisations restent claires). L'utilisateur peut aussi agrandir ou faire un zoom sur la visualisation d'un individu en particulier, changer l'ordre des dimensions et changer la méthode de visualisation, (figure 1-c). Les individus peuvent être affichés individuellement avec les matrices de scatter-plots (figure 1-d), les coordonnées parallèles (figure 1-e) ou avec la méthode Star Plot (figure 1-f). Trois possibilités sont offertes à l'utilisateur pour sélectionner une solution potentielle en cliquant directement sur la visualisation de l'individu, en le sélectionnant sur la partie droite de l'interface (figures 1-c, d, e, f)) ou en faisant un clic droit sur une visualisation en particulier, le choix de la sélectionner est alors offert à l'utilisateur (figure 1-b). Une fois les sélections effectuées les solutions apparaissent de couleurs différentes, comme sur l'exemple de la figure (1-a) où les solutions 1, 2, 3 et 4 sont sélectionnées. L'utilisateur doit cliquer sur l'onglet "Next" pour relancer l'AG sur quelques générations avant d'avoir d'autres visualisations.

3 Déroulement de l'algorithme

3.1 Algorithme Viz-IGA

Pour une taille du sous-ensemble de dimensions *s* fixée
1- Génération aléatoire de la population initiale
2- Vérifier les conditions de la population (pas de clones, pas de dimensions identiques dans un même individu)
3- Evaluation :
1- Première évaluation automatique des individus selon le critère à base de distance pour la détection d'outlier
2- Seconde évaluation visuelle par l'utilisateur toutes les 100 générations
4- Sélection :
1- Première sélection par l'utilisateur (soit E' = ensemble des solutions sélectionnées)
2- Seconde sélection par tournoi (on sélectionne aléatoirement deux individus, on ne garde que le meilleur)
5- Croisement mutation (en respectant conditions du point 2)
6- Si stagnation pendant η générations :
1- Muter quelques individus de E'
2- Générer de nouveaux individus
7- Aller à 2 ou fin

3.2 Description des étapes

Notre objectif principal est d'obtenir des visualisations de données significatives et pas trop surchargées, il est préférable de fixer s à de petites tailles (<10 pour que les visualisations restent claires). Nous avons choisi de représenter les données à l'aide des coordonnées parallèles (Inselberg, 1985) et des matrices de scatter-plot (Carr et al., 1987), cependant, ceci n'est pas figé, on pourra remplacer ou introduire d'autres méthodes de visualisation de données, nous avons notamment la méthode Star Plot (Card et al., 1999) dans les exemples présentés. Nous utilisons un algorithme génétique pour la recherche de sous-ensembles de dimensions pertinentes. L'interaction avec l'utilisateur intervient sur certaines générations afin de ne pas faire converger l'AG trop rapidement. Nous faisons intervenir l'utilisateur dans le processus de recherche dans deux étapes : l'évaluation et la sélection.

Population initiale : les individus de l'algorithme génétique représentent des sous-espaces de dimensions constitués à partir des dimensions décrivant l'ensemble des données. Une fois la population de départ prête, nous l'évaluons une première fois à l'aide d'un critère d'évaluation à base de distance décrit dans (Boudjeloud et Poulet, 2005).

Evaluation automatique : vu le nombre important de combinaisons de dimensions possibles, il est nécessaire de faire une sélection de dimensions en utilisant ce critère de validité des sous-espaces avant de les présenter à l'utilisateur. Une fois cette présélection automatique faite, l'utilisateur peut intervenir interactivement selon que la visualisation générée le satisfait ou pas.

Evaluation interactive : une fois la population évaluée et triée selon les différents objectifs, nous présentons à l'utilisateur 9 visualisations. Ces visualisations représentent la projection des données dans des sous-ensembles de dimensions choisis aléatoirement dans la population de l'AG (un individu de l'AG représente un sous-ensemble de dimensions, 9 individus sont choisi aléatoirement et sont présentés visuellement). Notre choix s'est fixé à 9 représentations pour ne pas surcharger l'interface. Les solutions sont représentées par des projections en coordonnées parallèles ou d'autres méthodes de visualisation selon le choix de l'utilisateur (figure 1-b). Nous opérons un croisement et une mutation, puis, toutes les 100 générations, nous proposons à l'utilisateur d'autres visualisations, il peut en sélectionner certaines s'il le souhaite en cliquant dessus, selon qu'elles sont assez significatives pour lui. Pendant ces 100 générations l'AG travaille tout seul sans intervention de l'utilisateur. L'algorithme prend en compte le choix de l'utilisateur pour les prochaines générations dans le processus de recherche de deux manières. Nous avons choisi 100 générations pour que l'AG puisse éliminer les solutions redondantes, les moins intéressantes automatiquement et éviter d'avoir toujours les mêmes solutions qui seront affichées (présentées à l'utilisateur).

Sélection interactive : les solutions sélectionnées par l'utilisateur seront stockées dans une mémoire E' que nous faisons intervenir dans deux étapes de l'algorithme, la première étant la reproduction, la seconde pour remédier à la stagnation de la recherche.

Reproduction : nous faisons intervenir les solutions de E' (sélectionnées par l'utilisateur) dans la reproduction de la façon suivante : chaque nouvel enfant généré aura une partie des gènes d'un parent issu de E' et une partie des gènes d'un parent issu d'une sélection par tournoi où l'on sélectionne aléatoirement et uniformément 2 individus en ne gardant que le meilleur.

Stagnation : dès que la solution stagne (ne s'améliore pas pendant un certain nombre de générations) nous générons de nouvelles solutions à partir des solutions de E' (sélectionnées par l'utilisateur) en les faisant intervenir dans le processus de mutation. Lorsqu'un gène doit être muté, il sera changé par un gène d'un individu de E' (l'emplacement du gène et l'individu de E' sont aléatoires).

L'espace de recherche étant grand, il est important d'avoir une grande capacité d'exploration. Ces mécanismes permettent de maintenir une diversité dans la population en introduisant à certains moments de nouveaux individus. Ils permettent aussi d'éviter une convergence prématurée ou une stagnation des solutions. Nous utilisons ces mécanismes lorsque le meilleur individu est le même durant un certain nombre de mutations η_{mut} et de croisements η_{croi} (ces deux paramètres sont exprimés en pourcentage de la taille de la population). Alors, tous les individus de la population qui ont une valeur d'évaluation en dessous de la moyenne de la population (sous la médiane) sont remplacés par de nouveaux individus générés en respectant les conditions de non-présence de clones et de dimensions identiques dans un même individu. La différence entre notre AGI et les autres AGIs existants est que nous faisons coopérer les deux méthodes visuelle et automatique et que nous faisons intervenir l'utilisateur dans deux processus de l'AG : l'évaluation et la sélection.

4 Résultats et interprétation

Le système a été implémenté sous Windows 2000 dans un environnement très intuitif pour l'utilisateur. Les différentes possibilités de Viz-IGA ont été testées sur des ensembles de données du Kent Ridge Biomedical Dataset Repository (Jinyan et Huiquing, 2002). Les

différentes figures (3a, 3b, 3c, 3d) sont créées à partir d'un exemple de détection d'outlier sur l'ensemble de données Breast cancer, Lung cancer, Ovarian et MLL. Notre méthode permet de retrouver des outliers sur des sous-espaces de dimensions identiques à ceux trouvés sur l'ensemble total des données détectés par LOCI (Papadimitriou et al., 2003) un algorithme récent, qui détecte les éléments outliers de l'ensemble des données que nous avons testé (Boudjeloud et Poulet, 2005). De plus, Viz-IGA permet de mettre en évidence les dimensions prépondérantes et souligner la pertinence et l'intérêt de certaines d'entre elles pour la détection d'outlier. Il permet d'isoler et de voir correctement l'élément outlier. Le nombre de dimensions peut être réduit jusqu'à un facteur 1000 en moyenne sans perte significative d'information, puisque nous arrivons à retrouver le même outlier dans les sous-espaces de dimensions que sur l'ensemble total des données. Il faut entre 2 et 10 minutes pour arriver à la visualisation la plus pertinente pour les problèmes mentionnés précédemment, soit environ de 4 à 20 interventions de l'utilisateur. Au-delà de 15 minutes, l'expérience montre que l'utilisateur se lasse et commence à se fatiguer. Une autre difficulté qui peut lasser l'utilisateur est la stagnation des solutions. Il peut en effet avoir plusieurs fois les mêmes solutions proposées. Il peut par exemple penser qu'il a obtenu la solution finale alors que c'est juste un optimum local. C'est le cas par exemple lors des tests effectués sur l'ensemble de données Colon où à chaque intervention de l'utilisateur on voit bien sur la figure 2 l'amélioration de la courbe et la convergence de l'algorithme en comparaison avec l'AG (Boudjeloud et Poulet, 2005) aux mêmes générations (Viz-IGA converge en moins de générations que l'AG). Cependant, on voit sur la courbe de VIz-IGA des moments de stagnation, par exemple entre les générations 800 et 1000, les générations 1200 et 1400 (il y a deux niveaux de stagnation) et les dernières générations à partir de la génération 1600. Dans ces cas là, il peut arriver que les mêmes solutions soient présentées à l'utilisateur plusieurs fois à la suite.

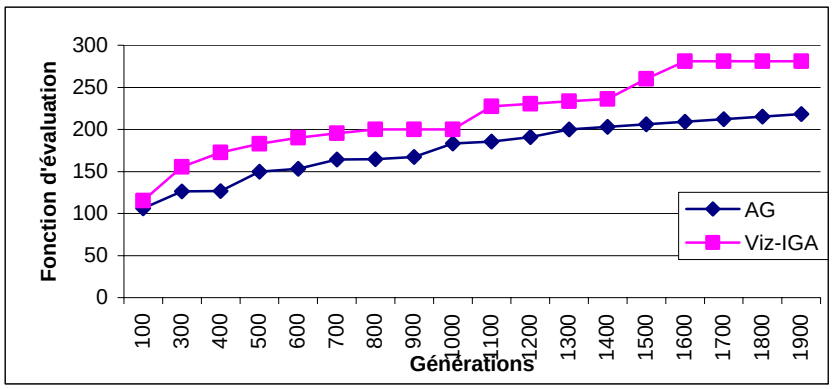
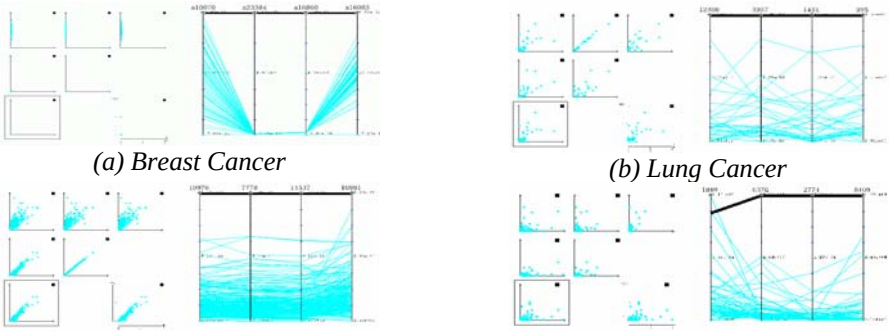


FIG. 2 – *Convergence de l'AG et de Viz-IGA.*

5 Modélisation de l'expertise

La visualisation des résultats obtenus sur quelques ensembles de données (figure 3) montre bien que les points détectés sont éloignés et présentent un comportement atypique par rapport au reste des données, néanmoins nous ne pouvons fournir plus d'explication sur le

type des points détectés par notre algorithme (par exemple erreur ou "outlier réel"). En effet, dans le cas de valeurs extrêmes on ne sait pas dire si cette valeur est une valeur possible ou non. Seul l'expert des données peut répondre à cette question. Dans le cas où le point détecté est une erreur on l'élimine de l'ensemble des données et dans le cas contraire on le garde dans les données car il peut représenter à lui seul des informations importantes. Un des moyens de combler cette lacune est de créer un modèle des données permettant de qualifier les éléments détectés comme outliers ou erreurs. Ainsi, étant donné un nouvel élément introduit dans l'ensemble des données, nous pourrons utiliser le modèle pour prédire son état : outlier, erreur ou donnée normale.



(a) Breast Cancer *(b) Lung Cancer*

(c) Ovarian *(d) MLL*

FIG. 3 – *Visualisation des résultats sur les différents ensembles de données.*

Nous proposons donc de construire un modèle de l'expertise de l'expert. Celui-ci doit tout d'abord étiqueter les éléments qui ont été détectés comme étant outliers (on peut supposer qu'il n'y a que 2 types d'éléments : les erreurs et les "vrais outliers"). A partir de cet ensemble de données étiquetées, on utilise un algorithme de classification supervisée (par exemple un algorithme d'induction d'arbre de décision) pour construire un modèle de l'expertise du spécialiste des données. Les nouveaux éléments outliers seront alors analysés avec le modèle construit et la présence de l'expert n'est plus indispensable pour qualifier ces outliers.

5.1 Construction du modèle

Concernant la partie expertise de la détection d'outlier, nous n'avons pas pu avoir accès à un ensemble de données avec un spécialiste pouvant étiqueter les éléments détectés. Nous avons donc décidé de le faire à partir de l'ensemble de données Colon Tumor (2000 dimensions, 62 éléments) de (Jinyan et al., 2002). Le nouvel ensemble de données est créé en rajoutant des éléments que nous avons étiqueté nous même d'erreur ou d'outlier. Par exemple des éléments qui présentent des valeurs extrêmes sont des erreurs et ceux qui présentent un comportement différent par rapport au reste des données sont des outliers. Nous obtenons donc l'ensemble Colon plus quelques nouveaux éléments. L'ensemble de données Colon a 62 éléments, 5 ont été détectés comme outlier par notre algorithme, nous les étiquetons comme tel, nous créons leurs clones (5), ces clones ont les mêmes valeurs que les originaux sur l'ensemble des dimensions mais sont présentés dans un ordre différant en permutant les valeurs de certaines dimensions. Nous rajoutons au nouvel ensemble de données 10 éléments avec plusieurs valeurs extrêmes qui vont être considérés comme erreurs. Nous obtenons un

ensemble de données que nous avons appelé "Colon-Bis" de 2000 dimensions, 77 éléments et trois classes :

Classe 1 : données correctes (57 éléments).
Classe 2 : outliers (10 éléments).
Classe 3 : erreurs (10 éléments).

L'ensemble de données crée n'a qu'un petit nombre d'éléments erreurs et d'éléments outliers, s'ils étaient nombreux, ils ne seraient plus considérés comme tels. Nous avons pris pour l'ensemble d'apprentissage 67 éléments choisis aléatoirement et les 10 éléments restants pour l'ensemble de test. Reste à choisir un algorithme d'apprentissage qui pourra prédire la classe des nouveaux individus. De nombreux algorithmes d'apprentissage automatique peuvent être utilisés, nous avons choisi comme algorithmes les k-PPV (Cover et Hart, 1967), C4.5 (Quinlan, 1993), CART (Breiman et al., 1984) et LibSVM (Fan et al., 2005). Nous avons effectué des tests dont nous présentons les résultats dans le tableau 1. Ces résultats sont très satisfaisants, nous arrivons à prédire les nouveaux éléments avec un taux de précision de 100% avec le modèle établi par LibSVM. Une fois le modèle établi, le besoin d'étiqueter par le spécialiste des données n'est plus nécessaire.

Algorithmes	Taux de bon classement (%)
LibSVM	100
CART	99
C4.5	98.5
k-PPV	90

TAB. 1 – *Résultats obtenus sur l'ensemble de données Colon Bis.*

6 Conclusion

Nous avons présenté un algorithme génétique semi-interactif pour la sélection de dimensions appliqué à la détection d'outlier. Nous avons introduit une nouvelle représentation de l'individu de l'algorithme génétique. Notre choix s'est fixé sur des petites tailles de sous-ensembles de dimensions pour faciliter l'interprétation visuelle des résultats et souligner la pertinence des dimensions pour chacune des applications, ajoutant ainsi un aspect d'aide à la décision. Cependant, l'utilisateur est libre de fixer la taille des sous-ensembles de dimensions. Il peut aussi intervenir sur le choix de la méthode visuelle utilisée et sur l'ordre des dimensions dans la visualisation proposée. Notre algorithme nous permet la détection d'outliers dans des ensembles de données ayant un grand nombre de dimensions en n'utilisant qu'un sous-ensemble de dimensions de l'ensemble initial. Puisque le nombre de dimensions utilisé est faible, on peut ensuite visualiser ces éléments (à l'aide de matrices de scatter-plot ou de coordonnées parallèles) pour permettre à l'utilisateur de choisir les solutions qui lui paraissent pertinentes. Elles sont alors utilisées pour générer et visualiser d'autres solutions et aussi pour expliquer et valider les résultats obtenus. Il ne faut pas oublier que l'on travaille sur des données de grandes dimensions. Cette étape n'est possible que parce que nous n'utilisons qu'un sous-ensemble restreint de dimensions de l'ensemble de données initial. Cette interprétation des résultats serait absolument impossible en considérant l'ensemble des dimensions comme le montre la figure 4.

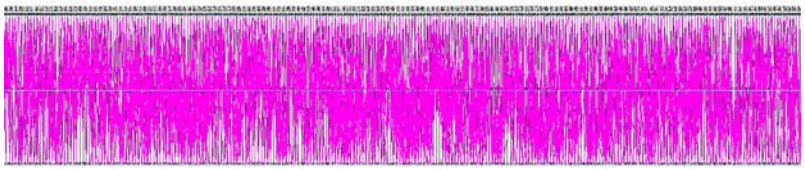


FIG. 4 – *Visualisation de quelques centaines d'attributs de l'ensemble de données Colon Tumor.*

Nous pensons étendre nos applications à des données symboliques et aussi améliorer notre méthode pour optimiser l'ordre des dimensions dans les visualisations en utilisant des critères de validité et étendre la méthode au clustering. Viz-IGA a permis de montrer que l'on peut gagner beaucoup en augmentant l'interaction entre l'expert du domaine et son outil de fouille de données. Pour finir, l'étude des AGI est certainement prometteuse dans d'autres domaines. Ces algorithmes proposent une interaction très simple et efficace pour un utilisateur non informaticien, ce qui peut leur assurer un certain succès dans les applications nécessitant une interaction homme/machine. Nous avons fait coopérer les méthodes automatiques et les méthodes de visualisation de données sur deux aspects, l'interaction avec l'utilisateur dans le processus de recherche en le faisant participer dans la sélection et l'évaluation des solutions proposées par l'AG et dans l'interprétation et la qualification des éléments détectés comme outlier à travers le modèle d'expertise du spécialiste des données. Nous avons proposé une partie expertise, à l'aide des visualisations présentées l'expert des données peut qualifier les outliers détectés (par exemple en deux classes : erreur ou élément significativement différent de la masse). Il ne faut pas oublier que l'on travaille sur des fichiers de grandes tailles. Cette étape n'est possible que parce que nous n'utilisons qu'un sous ensemble restreint de dimensions de l'ensemble de données initial. Cette qualification des outliers serait absolument impossible en considérant l'ensemble des dimensions comme l'illustre très bien l'exemple de la figure 4 où l'on ne peut détecter aucune information à propos des éléments de l'ensemble de données ou des dimensions. Une fois la qualification effectuée, nous utilisons un algorithme d'apprentissage pour créer un modèle de l'expertise du spécialiste des données. Les nouveaux outliers peuvent alors être qualifiés par le modèle construit sans la présence de l'expert des données. Cette étape souligne l'importance de la visualisation de données pour l'interprétation des résultats et son apport pour l'aide à la décision. Les tests effectués pour l'expertise ont été effectués sur un ensemble de données artificiel crée par nos soins car nous n'avons pas pu avoir accès à un ensemble de données et un spécialiste pouvant qualifier les éléments détectés d'erreur ou outlier réel. Nous avons obtenu des résultats satisfaisants par ce premier travail qui nous a permis de faire participer l'utilisateur dans le processus de recherche de sous-ensembles de dimensions pertinents pour détecter, interpréter visuellement et qualifier des éléments outliers.

Références

Boudjeloud, L. et F. Poulet (2005). Détection et interprétation visuelle d'outliers dans les grands ensembles de données. *Numéro spécial de la Revue des Nouvelles Technologie de l'Information : Visualisation et Extraction des Connaissances*, F. Poulet et P. Kuntz Eds, (à paraître).

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Card, S., J. MacKinlay and B. Shneiderman (1999). *Readings in information visualization : Using vision to think*. Morgan Kaufman.

Carr, D. B., R. J. Littlefield and W. L. Nicholson (1987). Scatter-plot matrix techniques for large n. *Journal of the American Statistical Association*, 82(398):424-436.

Cover, T. M. and P. E. Hart (1967). Nearest neighbor pattern classification. *In IEEE Transaction on information theory*. 13: 21-27.

Dash, M., H. Liu and J. Yao (1997). Dimensionality reduction for unsupervised data. *In Proceedings of 9th IEEE International conference on tools with artificial (ICTAI).*

Fan, R-E., P-H. Chen and C-J. Lin (2005). Working set selection using the second order information for training svm. *Technical report*, Department of Computer Science, National Taiwan University. Logiciel disponible en ligne (http ://www.csie.ntu.edu.tw/ cjlin/libsvm/) accédé en septembre 2005.

Hayashida, N. and H. Takagi (2000). Visualised IEC : Interactive evolutionary computation with multidimensional data visualization. *In Industrial electronics, control et instrumentation, IECON2000*, 2738-2743.

Inselberg, A. (1985). The plane with parallel coordinates. *In Special Issue on Computational Geometry*, 1:69-97.

Jinyan, L. and L. Huiqing (2002). Kent ridge bio-medical data set repository. http ://sdmc.-lit.org.sg/GEDatasets accédé en septembre 2005.

Jourdan, L. (2003). Métaheuristiques pour l'extraction des connaissances, application à la génomique. *Thèse de doctorat*, Université des Sciences et Technologies de Lille.

Narendra, P.M. and K. Fukunaga (1977). A branch and bound algorithm for feature subset selection. *In IEEE Transactions in Computers*, 26:914-922.

Papadimitriou, S., H.Kitawaga, P. B. Gibbons and C. Faloutsos (2003). LOCI: Fast Outlier Detection Using the Local Correlation Integral. *19th International Conference on Data Engineering, Sponsored by the IEEE Computer Society ICDE'03*.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Takagi, H. (2001). Interactive evolutionary computation : Fusion of the capacities of EC Optimization et human evaluation. *In Proceedings of the IEEE*, 89:1275-1296.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Summary

We present a semi-interactive genetic algorithm for dimensions selection in high dimensional data sets for outlier detection. Several data mining algorithms have problems with high dimensional data sets, one solution is to carry out a pre-processing in order to

retain only "interesting" dimensions. In addition we want to give a more important role to the user in the search process, for that we chose to use an interactive genetic algorithm. In this type of approach the user replaces the genetic algorithm fitness function. Our approach does not completely eliminate this function but we use the user evaluation with it, then we introduce a semi-interactive genetic algorithm. Finally, the important reduction of the dimensions number enables us to display the algorithm results of the outlier detection. This visualization allows the data expert to label the atypical elements, for example if they are errors or simply individuals different from the mass.

Tree-View : post-traitement interactif pour les arbres de décision

Nguyen-Khang Pham*, Thanh-Nghi Do*

*College of Information Technology, Cantho University
1 Ly Tu Trong street, Cantho City, Vietnam
pnkhang@cit.ctu.edu.vn
dtnghi@cit.ctu.edu.vn

Résumé. Nous présentons une nouvelle méthode graphique interactive, Tree-View permettant de visualiser les résultats obtenus par un algorithme d'apprentissage automatique d'arbres de décision comme C4.5. Le Tree-View représente sous forme graphique les résultats des algorithmes d'arbre de décision les rendant plus accessibles que des colonnes de chiffres ou un ensemble de règles habituellement en sortie de ces algorithmes. Il permet aussi à l'utilisateur interactivement d'extraire des règles inductives et d'élaguer l'arbre en post-traitement. L'intérêt de l'utilisation du Tree-View est de permettre à l'utilisateur de mieux comprendre les résultats de l'algorithme d'arbres de décision.

1 Introduction

Les chercheurs de l'université de Berkeley ont estimé que la quantité d'informations dans le monde augmente d'environ deux exa (10^{18}) octets tous les ans [Lyman et al., 2003]. Une telle masse d'informations est trop complexe pour pouvoir être appréhendée par un utilisateur. Le domaine de l'extraction de connaissances à partir de données (ECD) s'est développé pour répondre à cette volonté de découverte de connaissances. D'après [Fayyad et al., 1996], la définition de l'ECD est : « un processus non trivial d'identification de connaissances inconnues, valides, potentiellement exploitables et compréhensibles dans les données ». Le but de l'ECD est de pouvoir extraire des informations intéressantes contenues dans de grands ensembles de données pour une application connue a priori. L'intérêt des connaissances extraites est validé en fonction du but de l'application. Seul l'utilisateur peut déterminer la pertinence des résultats obtenus par rapport à ses objectifs. L'idée ici est d'augmenter la participation de l'utilisateur dans le processus d'ECD. L'utilisation des méthodes graphiques interactives [Keim, 2002] permet à l'utilisateur d'être plus impliqué dans le processus d'ECD.

Les techniques de visualisation de données apportent aussi leur contribution au processus d'ECD et peuvent intervenir à plusieurs niveaux.

En prétraitement de données, elles sont utilisées pour visualiser des données initiales. On peut détecter de manière visuelle des tendances, corrélations dans ces données grâce à la visualisation. Cette étape peut également guider l'utilisateur dans le choix des algorithmes de fouille ou de leurs paramètres.

En post-traitement des algorithmes automatiques de fouille de données, les méthodes de visualisation [Poulet, 2001a] sont utilisées pour interpréter et évaluer les résultats en se basant sur des représentations graphiques plus accessibles que des colonnes de chiffres ou un ensemble de règles.

De nouvelles méthodes [Ankerst, 2000], [Poulet, 2001b] remplacent l'algorithme automatique de fouille par un algorithme graphique interactif de fouille visuelle de données. Cela apporte au moins les avantages suivants : on utilise les compétences et connaissances du spécialiste du domaine des données lors de la construction du modèle, on profite des capacités humaines en reconnaissance de formes, l'utilisateur a une meilleure compréhension du modèle qu'il construit et une plus grande confiance dans ce modèle.

Nous présentons une nouvelle méthode graphique interactive, Tree-View, permettant à l'utilisateur d'être plus impliqué dans le post-traitement des algorithmes d'arbres de décision comme C4.5 [Quinlan, 1993]. Le Tree-View permet de représenter de grands arbres sous forme graphique. Il permet à l'utilisateur d'exploiter de manière interactive l'arbre obtenu et de faciliter l'extraction des règles inductives. Le Tree-View fournit des bons outils pour aider l'utilisateur à élaguer lui-même l'arbre en post-traitement. L'intérêt de l'utilisation du Tree-View est de permettre à l'utilisateur de mieux comprendre les résultats de l'algorithme d'arbres de décision. Nous avons fait des évalué le Tree-View en se basant sur les ensembles de données de Statlog [Michie et al., 1994].

Le paragraphe 2 présente le principe du Tree-View en post-traitement de l'algorithme d'arbres de décision C4.5. Un exemple du Tree-View est présenté dans le paragraphe 3 avant de conclure sur nos travaux.

2 Tree-View

Un arbre de décision C4.5 est une représentation graphique d'une procédure de classification où le noeud interne de l'arbre est le nœud de décision, il lui est associé un test sur une dimension, la feuille est la classe. A partir de l'ensemble d'apprentissage, on essaie de construire l'arbre de décision en commençant à la racine de l'arbre, en descendant, à chaque étape on cherche une dimension qui permette de bien séparer une classe des autres classes de l'ensemble d'apprentissage, on va continuer à construire récursivement l'arbre jusqu'aux feuilles (classes). Pour éviter de construire un arbre trop grand possédant souvent beaucoup de feuilles avec très peu d'exemples, on essaie d'élaguer l'arbre obtenu lors de la phase de construction. L'algorithme d'arbres de décision fournit des méthodes efficaces qui obtiennent de bons résultats dans la pratique (voir http://www.kdnuggets.com/polls/2004-/deployed_data_mining_techniques.htm). Les arbres de décision sont compréhensibles par tout utilisateur (si leur taille est raisonnable). Les règles de décision sont sur les chemins menant de la racine aux feuilles.

En sortie de l'algorithme C4.5, on obtient des résultats sous forme textuelle sans aucun autre moyen permettant à l'utilisateur d'exploiter les modèles obtenus. Il n'est pas toujours facile d'interpréter et même de comprendre un arbre de grande taille (100 noeuds). La compréhensibilité du modèle est aussi importante que sa précision même si elle n'est pratiquement jamais évaluée dans les algorithmes de fouille de données. L'amélioration de la compréhensibilité peut aider l'utilisateur à mieux comprendre et utiliser le modèle.

Nous présentons une nouvelle méthode graphique interactive, Tree-View, permettant à l'utilisateur d'être plus impliqué dans le post-traitement de l'algorithme d'arbres de décision.

Le Tree-View permet une visualisation interactive des résultats obtenus par l'algorithme d'arbres de décision. Un nœud de l'arbre est représenté par un noeud graphique dont la couleur correspond à la classe d'une feuille ou la classe majoritaire du nœud interne. L'état d'un noeud concernant le nombre d'erreurs et d'individus est associé au noeud. En cliquant avec la souris sur le noeud, on descend d'un niveau dans l'arbre en affichant les noeuds correspondant aux fils du nœud courant. La figure 1 est la visualisation de l'arbre de décision des données Shuttle (43500 individus, 9 dimensions, 7 classes).

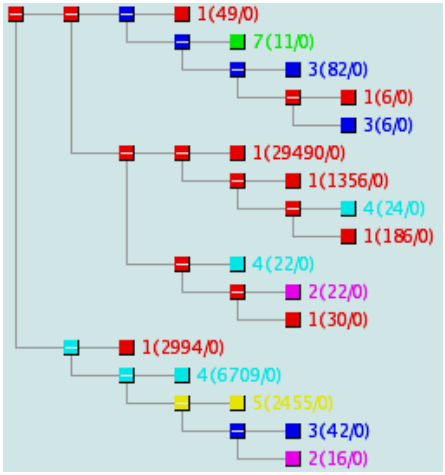


FIG. 1 – *Visualisation de l'arbre de décision des données Shuttle*

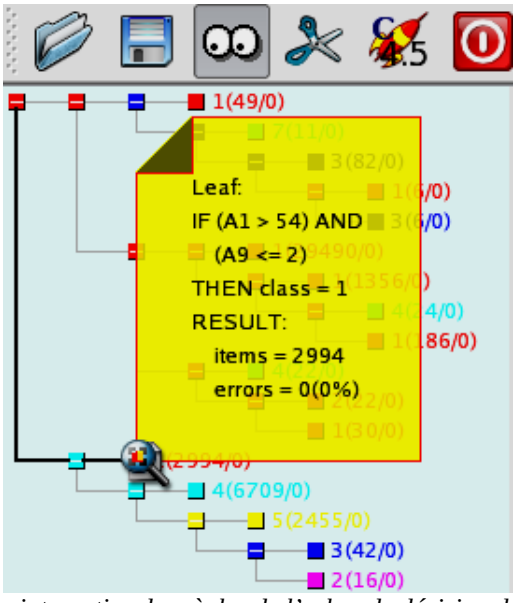


FIG. 2 – *Extraction interactive des règles de l'arbre de décision des données Shuttle*

L'utilisateur peut visualiser l'information des coupes. Un outil graphique permet à l'utilisateur d'extraire de manière intuitive des règles de l'arbre. Le chemin en gras menant de la racine au noeud courant est une règle de décision comme sur la figure 2. Dans le cas où la taille des arbres (la profondeur, le nombre de nœuds) est grande, le Tree-View peut réduire l'arbre pour avoir une vue globale et descendre dans le sous arbre pour une vue locale.

Nous avons proposé une nouvelle vue d'arbre de grande taille. Nous utilisons le panel ellipsoïde pour représenter un arbre. La racine est au centre du panel. On peut tourner l'arbre d'après le grand diamètre pour facilement exploiter des informations des noeuds. De plus, nous utilisons aussi l'animation pour se focaliser sur le noeud courant. Les différentes échelles permettent d'avoir simultanément une vue globale et locale dans l'arbre. La figure 3 est un exemple de visualisation de l'arbre de décision des données SatImage (6435 individus, 36 dimensions, 6 classes).

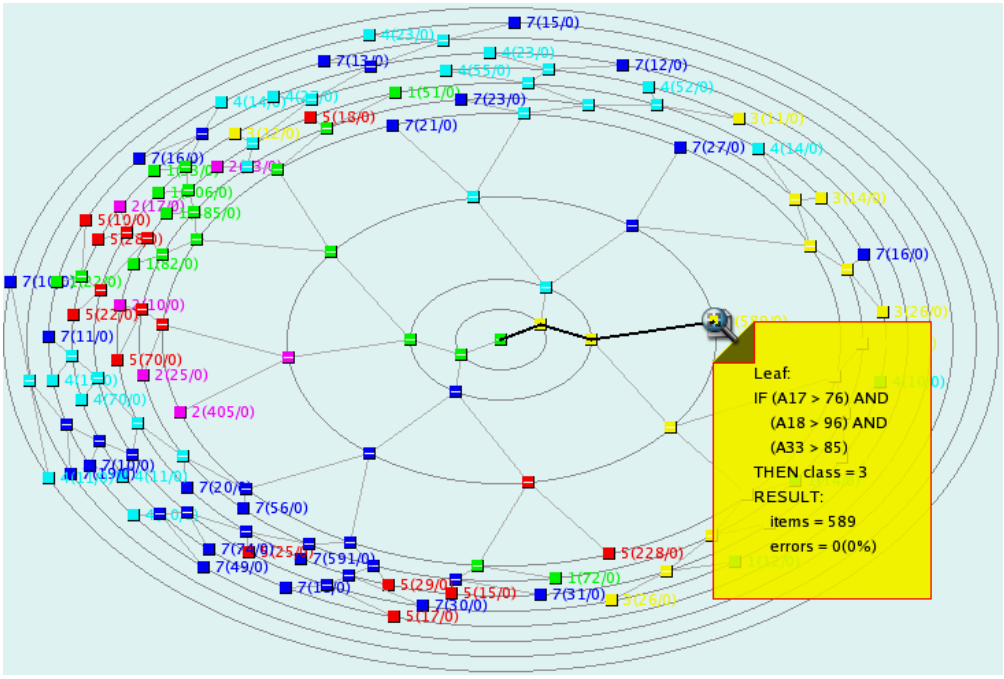


FIG. 3 – *Visualisation et extraction des règles de l'arbre de décision des données SatImage*

Le Tree-View permet de facilement exploiter les résultats obtenus par l'algorithme d'arbres de décision. L'utilisateur peut visualiser les résultats et extraire intuitivement des règles de décision. Le Tree-View fournit simultanément une vue globale et locale dans l'arbre.

L'utilisateur a aussi la possibilité d'être plus impliqué dans le post-traitement de l'algorithme C4.5. Le Tree-View fournit de bons outils permettant à l'utilisateur d'élaguer un arbre obtenu en sortie de l'algorithme C4.5. L'utilisateur se base sur la couleur et le nombre d'erreurs et d'individus du nœud courant pour pouvoir décider d'élaguer l'arbre comme sur la figure 4. Le Tree-View prend en compte cette modification dans le résultat.

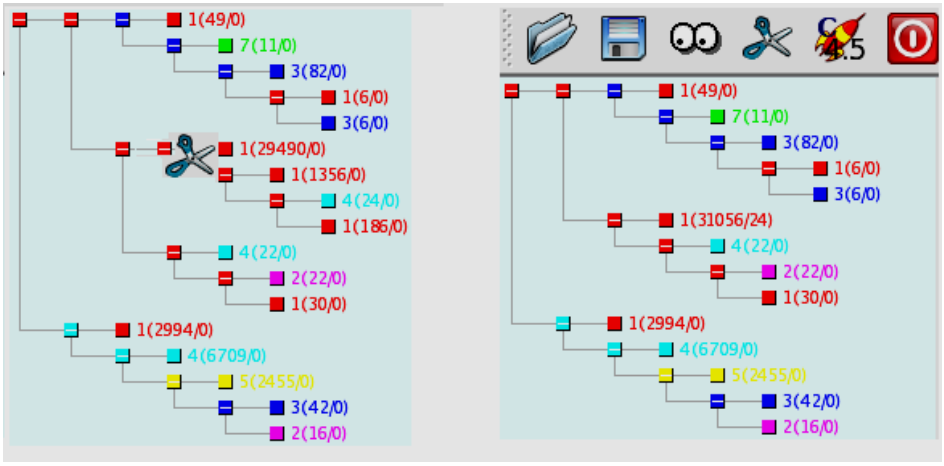


FIG. 4 – *Elagage de l'arbre de décision des données Shuttle*

3 Tree-View pour l'arbre de décision des données Segment

L'ensemble du programme est écrit en C/C++ sous Linux (PC), nous avons utilisé la librairie Qt [Trolltech, 2005] pour implémenter le Tree-View. Nous avons également intégré l'algorithme d'arbres de décision C4.5 dans l'outil. Les arbres de décision sont sauvegardés en format XML pour être facilement traités par des autres outils.

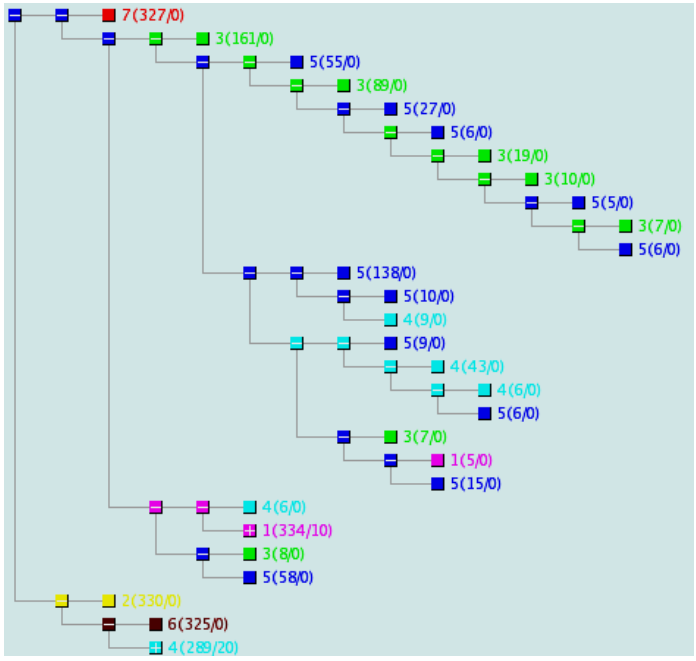


FIG. 5 – *Visualisation de l'arbre de décision des données Segment*

Tree-View

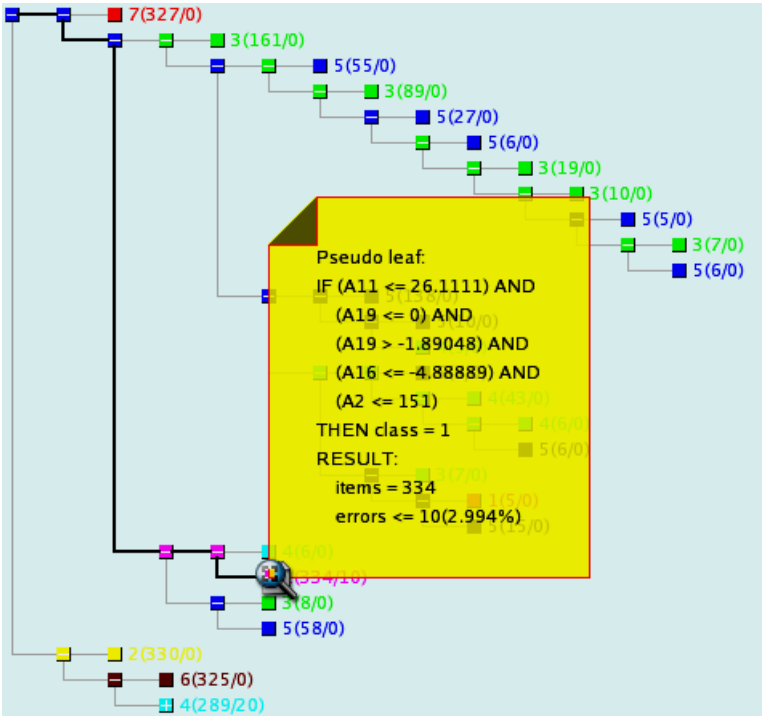


FIG. 6 – *Extraction interactive des règles de l'arbre de décision des données Segment*

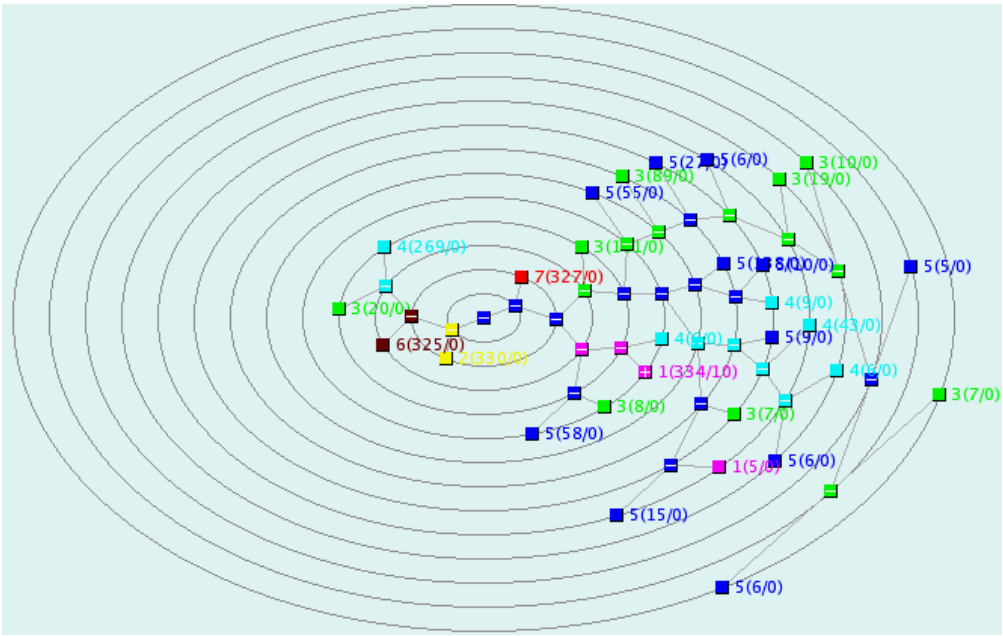


FIG. 7 – *Visualisation de l'arbre de décision des données Segment avec le panel ellipsoïde*

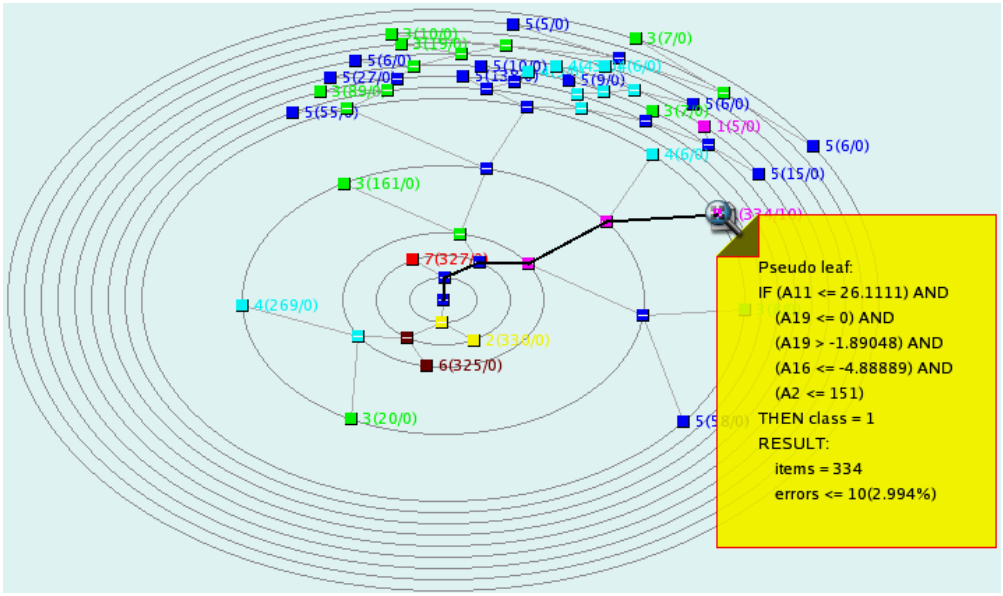


FIG. 8 – *Extraction interactive des règles de l'arbre de décision des données Segment avec le panel ellipsoïde*

Nous nous intéressons ici à l'utilisation du Tree-View pour visualiser un grand arbre de décision des données Segment (2310 individus, 19 dimensions, 7 classes). Le Tree-View visualise l'arbre de décision en sortie de l'algorithme C4.5 comme sur les figures 5 et 7. L'utilisateur extrait intuitivement des règles de décision (figures 6 et 8) et peut aussi élaguer l'arbre. Le Tree-View fournit simultanément une vue globale en réduisant l'arbre et locale en descendant dans l'arbre.

4 Conclusion et perspectives

Nous présentons une nouvelle méthode graphique interactive, Tree-View permettant à l'utilisateur d'être plus impliqué dans le post-traitement des algorithmes d'arbres de décision comme C4.5. Il peut aider l'utilisateur à mieux comprendre et utiliser les arbres de décision. Le Tree-View permet de représenter de grands arbres sous forme graphique. Il permet à l'utilisateur d'exploiter de manière interactive l'arbre obtenu et de faciliter l'extraction de règles inductives. Le Tree-View fournit de bons outils pour aider l'utilisateur à élaguer lui-même l'arbre en post-traitement. L'intérêt de l'utilisation du Tree-View est de permettre à l'utilisateur de mieux comprendre les résultats des algorithmes d'arbres de décision.

L'extension la plus immédiate de ce travail est d'utiliser un ensemble méthodes de visualisation pour mieux interpréter les grands arbres de décision.

Références

[Ankerst et al., 2000] M. Ankerst, M. Ester, et H-P. Kriegel. Towards an Effective Cooperation of the Computer and the User for Classification. Proceeding of KDD'00, 6th ACM SIGKDD, Boston, USA, 2000, pp. 179-188.

[Fayyad et al., 1996] U. Fayyad, G. Piatetsky-Shapiro, et P. Smyth. From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), pp. 37-54, 1996.

[Keim, 2002] D. Keim. Information Visualization and Visual Data Mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), pp. 1-8, 2002.

[Lyman et al., 2003] P. Lyman, H. R. Varian, K. Swearingen , P. Charles, N. Good, L. Jordan et J. Pal. How Much Information, 2003. http://www.sims.berkeley.edu/research/-projects/how-much-info-2003/.

[Michie et al., 1994] D. Michie, D.J. Spiegelhalter et C.C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, 1994.

[Poulet, 2001a] F. Poulet. CubeVis: Voir pour Mieux Comprendre. Actes de SFDS'01, XXXIIIe Journées de Statistiques, Nantes, 2001, pp.637-640.

[Poulet, 2001b] F. Poulet. Construction Interactive d'Arbres de Décision. Actes de SFC'01, VIIIe Rencontres de la Société Francophone de Classification, Pointe à Pitre, 2001, pp. 275-282.

[Quinlan, 1993] J. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.

[Trolltech, 2005] Trolltech Inc. Qt4. 2005. http://www.trolltech.com/products/qt/index.html.

Summary

We present a new interactive graphical method, Tree-View for visualizing the results obtained by decision tree induction algorithms. Tree-View represents large trees in a graphical mode more intuitive than the columns of numbers or the rule sets in output of usual algorithms. The user can easily extract inductive rules and prune the tree in a post-processing stage. The user has a better understanding of the obtained model with Tree-View.