Un aperçu de la fouille visuelle de données

Hanene Azzag*, David Da Costa**
Christiane Guinot***, **, Gilles Venturini**

*Laboratoire d'Informatique de Paris-Nord
99 Avenue J-B. Clément, 93430 Villetaneuse.
hanene.azzag@lipn.univ-paris13.fr
**Laboratoire d'Informatique, Université François-Rabelais de Tours,
64, Avenue Jean Portalis, 37200 Tours.
david.dacosta@univ-tours.fr, venturini@univ-tours.fr
***CE.R.I.E.S., 20 rue Victor Noir, 92521 Neuilly-sur-Seine Cedex
christiane.guinot@ceries-lab.com

Résumé. Nous présentons dans cet article un aperçu de la fouille visuelle de données. Pour commencer, nous situons ce domaine par rapport à d'autres approches et nous en rappelons les principes fondateurs. Ensuite, nous montrons qu'il existe de nombreux points de vue pour aborder les travaux en fouille visuelle de données : les données ou connaissances à visualiser, la tâche à accomplir, la représentation visuelle choisie, la méthode de calcul de cette représentation ou encore le domaine d'application traité. Nous choisissons tout d'abord le point de vue des données à visualiser en détaillant des approches représentatives pour la visualisation de données numériques, de données hiérarchiques et de documents. Ensuite, notre prenons le point de vue de la représentation visuelle choisie en présentant le domaine des métaphores visuelles utilisées pour la fouille de données. Nous finissons en traitant d'un domaine thématique particulier, l'analyse d'audience d'un site Web, et en concluant sur les perspectives en fouille visuelle de données.

1. Introduction

La base des Iris de Fisher (Fisher 1936) est un ensemble de données bien connu maintenant (Blake et Merz 1998) composé de 150 fleurs décrites par 4 caractéristiques numériques et un attribut de classes (3 classes possible). Imaginons que l'expert du domaine souhaite comprendre comment cette base est structurée et en particulier s'il existe du bruit, des groupes de points distincts ayant certaines caractéristiques, etc. Une première approche possible pour résoudre ce type de problème consiste à utiliser des outils d'apprentissage artificiel qui vont extraire par exemple pour chaque classe, des domaines de valeurs qui font que chaque groupe appartient à telle ou telle classe. Un avantage incontestable de ce type de méthode, si l'on omet les étapes souvent cruciales de réglages des paramètres, est d'être entièrement automatique et de ne pas faire appel à l'expert du domaine dans le processus de découverte des connaissances. Cependant, on peut aussi donner des inconvénients de ce type de méthodes : d'une part la compréhensibilité des résultats n'est pas toujours évidente pour l'expert, que cela soit en terme de représentation des hypothèses apprises (langage de représentation nécessitant quel investissement de la part de l'expert) ou d'explications sur la

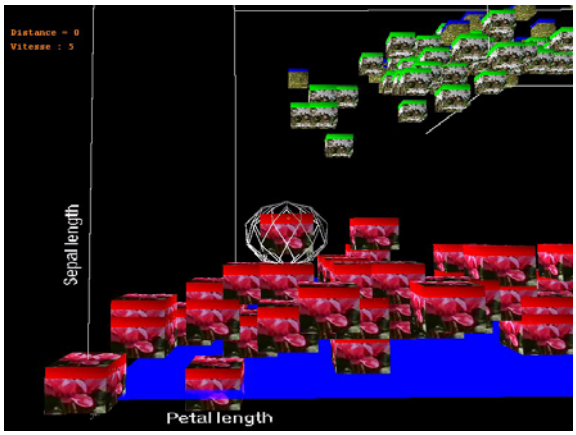


Fig. 1 – Visualisation de la base des Iris de Fisher en utilisant des photos de fleurs pour représenter chaque donnée (Azzag et al. 2006).

manière dont la méthode fonctionne. D'autre part, il est illusoire de dire que l'expert n'interviendra pas dans un processus de fouille de données, ne serait-ce que du fait que seul l'expert final peut en réalité valider les résultats obtenus. Un deuxième type d'approches, celui de la fouille visuelle de données, va utiliser une philosophie complètement différente (Wong et Bergeron 1997) : il va s'agir d'une part de représenter les données sous la forme de signes visuels permettant une interprétation directe facilitant l'accomplissement de la tâche à réaliser par l'expert, et d'autre part de permettre une interaction, souvent très intuitive, entre l'expert et cette visualisation, prenant la forme généralement de « requêtes graphiques ». Cette approche vise donc à faire intervenir plus directement l'expert du domaine et tend à permettre une analyse rapide des données car utilisant les capacités d'interprétation visuelle de l'esprit humain. Mais il arrive aussi que cette distinction entre ces deux approches ne soit pas seulement philosophique, mais corresponde à une limitation des approches non visuelles. Par exemple, si l'on considère que les Iris sont également représentées par des photos (voir figure 1), comment aider l'expert à établir des corrélations entre ces photos et les autres attributs présents dans la base ? L'approche de la fouille visuelle de données peut répondre également à ce type de problématique en proposant une perception simultanée de nombreuses informations. Son principal inconvénient, outre le fait de ne pas être automatique, est de devoir trouver le couple représentation visuelle/interaction le plus adapté aux besoins de l'expert.

De nombreux auteurs ont tenté de modéliser le processus de visualisation et de fouille visuelle de données ou encore de décrire les travaux de ce domaine selon différents points de vue (voir par exemple (Leung et Apperley 1994), (Chi et Riedl 1998) ou encore (Bruley 1999)). On peut définir d'une manière générale un processus servant de canevas commun aux méthodes de fouille visuelle de données de la façon suivante :

1. Recueil des données brutes,
2. Normalisation des données,
3. Codage des données sous la forme de signes visuels : c'est la première étape clé dans la fouille visuelle de données. Il s'agit de définir comment les données

seront représentées visuellement à partir des valeurs prises par les attributs (ou prédicats),

4. Algorithme de « calcul » de la visualisation : dans certaines méthodes, cette étape est très réduite et consiste simplement à transcrire les valeurs des attributs en signes visuels. D'autres méthodes ont des complexités plus élevées, comme par exemple l'analyse en composantes principales ou le MDS (Green et al. 1989). Enfin, pour d'autres méthodes et problématiques, ce calcul peut correspondre à des problèmes NP-Difficiles par exemple, comme dans l'affichage de graphes,
5. Rendu graphique : cette étape est liée à la configuration de la machine et plus généralement aux aspects matériels (réalité virtuelle en particulier),
6. Interaction entre la visualisation et l'expert du domaine : il s'agit là encore d'une étape clé dans laquelle l'expert peut, par une approche de type « essai » et « erreur », explorer les données, vérifier des hypothèses ou découvrir des connaissances.

Ces différentes étapes montrent qu'il est possible d'aborder la fouille visuelle de données selon de nombreux points de vue, et de classer les différentes méthodes à l'aide de plusieurs caractéristiques précises.

Tout d'abord, en 1) le point de vue des données d'entrée consiste à savoir quel type de données peuvent être transformées en signes visuels par la méthode considérée : par exemple, il peut s'agir de données numériques ou symboliques, de séquences et autres données temporelles, mais aussi de textes, d'images ou bien d'autres données multimédia, ou encore des hiérarchies, des graphes ou des connaissances telles que des arbres de décision ou des règles d'association. Si l'on s'intéresse aux signes visuels utilisés pour représenter les données, alors il est possible de distinguer les représentations 1D, 2D ou 3D, les méthodes utilisant des icônes ou des métaphores. On peut également s'intéresser en 2) sur le point de vue des algorithmes utilisés pour calculer la visualisation, à l'instar de ce que l'on trouve dans le domaine de la visualisation de graphes : ces algorithmes peuvent être déterministes, stochastiques, heuristiques. On peut adopter aussi le point de vue général de la tâche à accomplir : est-ce qu'il s'agit d'effectuer des pré-traitements sur les données, de découvrir des connaissances ou de visualiser les connaissances produites par une autre méthode ? Dans ce cadre, on peut alors définir des tâches plus précises comme identifier une donnée, des groupes de données, des corrélations entre attributs, du bruit, des données aberrantes, etc. Enfin, en 3) il est possible également de classer les méthodes suivant les domaines d'applications visés (finance, Web, médecine, etc) ou encore selon le point de vue de l'interaction avec l'utilisateur avec la définition de requêtes graphiques, un point qui est très en relation avec le type de tâche à accomplir, ou encore l'utilisation de matériel spécifique tel que la réalité virtuelle.

Nous avons choisi dans cet article de nous focaliser sur quelques uns de ces points de vue en choisissant quelques exemples représentatifs de systèmes et méthodes de fouille visuelle de données. Ainsi, la section 2 se focalise sur le point de vue des données numériques (et symboliques), hiérarchiques ou textuelles. Ensuite, nous illustrons dans la section 3 le point de vue des attributs visuels en décrivant quelques exemples de métaphores. Dans la section 4, nous prenons le point de vue des applications en décrivant des utilisations de la fouille visuelle de données en analyse d'audience des sites Web.

2. Point de vue des données d'entrée : quelques exemples

2.1 Représentation de données en attributs/valeurs

Les données de types attributs/valeurs font ainsi partie des cas les plus largement traités, et spécialement les données numériques/quantitatives. De nombreuses méthodes existent pour visualiser et explorer de telles données. Historiquement, les *visages de Chernoff* (Chernoff 1973) sont une très bonne illustration des principes de visualisation de la fouille visuelle : les données sont représentées par des visages dont les caractéristiques dépendent des attributs. Ces caractéristiques sont la forme du visage, des yeux, du nez etc, mais aussi la position en 2D dans le graphique final. Ainsi, cette méthode exploite notre capacité à percevoir des visages, et donne la possibilité de détecter des corrélations entre les attributs notamment au sein de groupes de données. Elle est d'un apprentissage très facile pour l'utilisateur, mais elle est cependant limitée à quelques dimensions et surtout à un petit nombre de données puisque la perception correcte des visages nécessite qu'il n'y ait aucun chevauchement entre eux.

Les *« stick icons »* (Pickett et Grinstein 1988) codent des données numériques sous la forme d'une figure élémentaire composée de segments dont les angles vont dépendre des attributs (numériques). De la même manière que précédemment, le positionnement en 2D sur l'écran dépend de deux autres attributs. Cette méthode permet également de détecter des corrélations entre attributs, mais elle peut cette fois s'appliquer sur plusieurs milliers de données.

Ces deux approches ne proposent cependant que peu d'interaction avec les données. Au contraire, les *« scatter plots »* vont proposer une interaction originale par rapport à ces deux méthodes (Becker et Cleveland 1987). Il s'agit de visualiser tous les couples d'attributs sous la forme d'un ensemble de graphiques 2D, disposés à l'écran selon la matrice triangulaire supérieure ainsi considérée (couples d'attributs). La demi-matrice qui est restée vide sert alors à visualiser, en plus grand, le graphique sélectionné. Lorsque l'utilisateur sélectionne des points dans l'un des graphiques, il peut voir interactivement ces mêmes points dans les autres graphiques : si ces points forment un groupe dans un couple d'axes, alors on peut ainsi détecter si cette propriété se vérifie dans d'autres axes.

Les *coordonnées parallèles* (Inselberg 1985) vont représenter toutes les dimensions selon des barres parallèles verticales : une donnée est alors représentée par une ligne brisée. Cette méthode peut donc représenter un très grand nombre de données par des effets visuels de superposition et d'intensité lumineuse (plusieurs centaines de milliers de données numériques dans (Fua et al. 1999)). Des possibilités interactives existent et permettent de sélectionner des données sur un axe et d'observer alors seulement ces données.

Les *cartes de Kohonen* (Kohonen 1989) sont un exemple très connu de calcul conjoint d'une classification et d'une visualisation des données. En ce sens, cette approche se distingue nettement des précédentes car elle produit une classification : la contrepartie est de devoir laisser le temps de calcul nécessaire pour obtenir le résultat qui se trouve directement dans une forme visualisable. Le principe est de présenter successivement les données d'entrée à une carte neuronale qui va évoluer de manière à regrouper les données similaires sur les mêmes neurones. Ainsi, des neurones proches sur la carte représenteront des classes proches les unes des autres. Une application particulière a été développée sur des textes (Kohonen 1998) et permet de visualiser plusieurs dizaines de milliers de documents. Des possibilités d'interaction existent naturellement avec cette carte : sélection, zoom, distorsion.

Nous terminons cette section en présentant une méthode qui confirme toute la diversité des approches possible pour représenter ce type de données. *Radviz* (Korfhage 1991) permet de placer sur un cercle des attributs des données. Ensuite, chaque donnée est représentée au centre du cercle en fonction du poids de l'attribut dans cette donnée. Les données viennent donc se positionner comme si elles étaient suspendues au cercle par des ressorts dont la force dépend des poids. Cette approche à base de points d'intérêt permet de visualiser de grands ensembles de données (jusqu'à 1 million dans (Da Dacosta et Venturini 2006)).

Notons que des représentations existent pour traiter le cas des données temporelles (séries numériques par exemple) mais nous ne l'abordons pas ici (le lecteur intéressé peut se renseigner par exemple sur des systèmes comme *ThemeRiver* (Havre et al. 1999)).

2.2 Représentation de hiérarchies

Le cas des arbres et autres données organisées dans une hiérarchie est particulièrement important en fouille visuelle de données. L'exploration de telles structures est un problème courant car de multiples données et connaissances ont ce format (treillis, ontologies, arbres de décision, etc).

Pour permettre l'exploration visuelle et l'extraction de connaissances à partir de données hiérarchique, il est nécessaire d'employer un algorithme d'affichage d'arbres. D'un point de vue général, on peut distinguer par exemple l'affichage niveau par niveau (Janey 1992), de l'affichage radial (tous les nœuds d'un niveau donné sont sur un même cercle, tous les cercles ont le même centre) ou encore de l'affichage en ballon (les cercles ne sont plus concentriques). Ces différents types d'affichage ont été ensuite utilisés dans de nombreux systèmes de fouille visuelle de données hiérarchiques.

FSN (Tesler et Strasnick 1992) est un exemple de visualisation d'arbres (arborescence de fichiers) qui peut être comparé dans ses principes à l'affichage d'arbres de décision de *MineSet* (Brunk et al. 1997) (Thearling et al. 1998). Il s'agit de représenter l'arbre niveau par niveau, avec un pavé pour chaque nœud, ainsi que des colonnes positionnées sur ce pavé. Le pavé donne des informations générales sur le nœud (nombre de fichiers pour FSN, nombre de données pour MineSet). Les colonnes renseignent plus précisément l'utilisateur (taille de chacun des fichiers contenus pour FSN, proportion de chaque classe pour MineSet). Des couleurs peuvent également être utilisées (age du fichier pour FSN, pureté pour MineSet).

Un autre exemple bien connu de ce type de méthodes sont les *« cone trees »* (Robertson et al. 1991). Cette représentation utilise une vraie visualisation 3D dans laquelle un arbre est représenté par un cône qui contient l'ensemble de l'arbre. Le sommet du cône est la racine de l'arbre. La visualisation en 3D est interactive : en cliquant sur un nœud, on peut faire tourner l'arbre (certains sous-arbres) de manière à placer ce nœud face à l'utilisateur. Cette représentation a été utilisée notamment pour l'affichage de documents.

Les *arbres hyperboliques* développés par Inxight (Inxight 1999) utilisent une méthode particulière de distorsion afin de représenter l'ensemble de l'arbre (contexte) tout en zoomant sur une partie de celui-ci (focus), ce qui n'est pas le cas des méthodes présentées précédemment. Le principe utilisé est celui d'une représentation radiale mais projetée sur un disque 2D (ou une sphere 3D dans (Munzner et Burchard 1995)) à l'aide de la géométrie hyperbolique : les bords internes du cercle permettent, d'un point de vue mathématique, de représenter des rayons infinis et de garder ainsi l'ensemble de l'arbre à l'écran.

Nous mentionnons maintenant des approches de représentations d'arbres très différentes des précédentes, et qui ne vont plus s'appuyer sur la notion d'arc entre nœud. Le premier exemple est le *« Treemap »* (Shneiderman et al. 2000) (voir figure 2) qui représente un arbre

dans un ensemble imbriqué de rectangles. Ainsi, le nœud racine représente le rectangle initial qui englobera tous les autres. Ensuite, ce rectangle est découpé en autant de sous-rectangles qu'il y a de fils au nœud considéré, et ainsi de suite pour chacun des sous-arbres. Le sens de découpe est alterné d'un niveau à l'autre. La taille des rectangles dépend de l'importance du nœud (nombre de fichiers pour un système de gestion de fichier, nombre de données pour un arbre de décision), et l'on peut utiliser différentes couleurs pour chaque rectangle (type de fichiers, etc). Cette approche a été utilisée notamment pour visualiser un arbre à 1 million de nœuds (Fekete et Plaisant 2002). Des possibilités interactives existent également pour se focaliser sur une partie de l'arbre (mais généralement, il y a perte du contexte). Nous mentionnons également que d'autres exemples de visualisation d'arbres de décision existent comme (Ankerst et al. 1999).

2.3 Représentation de documents

Un autre cas particulier important de types de données est le texte. Les études en représentation et fouille visuelle de documents sont principalement motivées par la recherche d'information dans les bases documentaires et Internet (voir un survol dans (Mokaddem et al. 2006)). Des systèmes existent cependant en dehors de ce cadre, comme *SeeSoft* qui permet de visualiser sur un seul écran des dizaines de milliers de lignes de code (Eick et al. 1992) avec des possibilités de zoom et d'interaction (affichage en couleur selon certains critères, comme le niveau d'imbrication des boucles par exemple). Des systèmes similaires lui ressemblent pour la recherche d'information : dans l'interface *J24* (Ogden et al., 1998), chaque document est représenté par une colonne dans laquelle peuvent apparaître en couleur des mots clés sélectionnés par l'utilisateur. De même, une vue sous forme de liste classique est disponible ainsi qu'un zoom sur le document sélectionné.

Des méthodes particulières ont donc été développées pour représenter l'information textuelle. Tout d'abord, parmi les systèmes précédemment cités, nous pouvons mentionner les cartes de Kohonen (le système *ET-Map* permet la représentation de plus de 100.000 pages Web), les « cone trees » ou encore les « Treemap »**.**

Les *« Tile Bars »* (Hearst 1995) représentent une avancée importante puisqu'elles vont permettre de visualiser de nombreuses informations sur les documents : chaque document est une barre dont la longueur est proportionnelle à la longueur du document et dont le contenu est représenté par des carrés colorés indiquant la proportion de mots-clés dans chaque partie du texte.

Radial (Au et al. 2000) (voir figure 2) utilise des principes similaires à Radviz (voir section précédente) mais en considérant cette fois que les points d'intérêts placés sur le cercle sont des mots-clés. Chaque document se positionne alors en fonction de son contenu et de la proportion de ces mots-clés. Des opérations interactives sont possible et offrent des fonctionnalités intéressantes à l'utilisateur : en cliquant sur un document, on peut faire apparaître les mots-clés les plus significatifs sur le cercle. Inversement, en cliquant sur un mot-clé du cercle, on peut faire apparaître les documents qui lui sont le plus pertinents. Ces mots-clés peuvent également être déplacés hors du cercle. *SQWID* (McCrickard et Kehoe 1997) (voir figure 2) utilise également le principe de positionnement par ressorts (vis à vis de trois mots-clés) mais en ajoutant cette fois des déplacements liés à la configuration locale (des documents ne doivent pas être trop proches les uns des autres par exemple) ainsi qu'une visualisation avec des codes de couleurs indiquant la présence des mots-clés (ou le nombre de pages contenues dans le site).

D'autres systèmes peuvent être mentionnés comme par exemple *Envision* (Nowell et al. 1996), qui représente un ensemble de documents sous la forme d'un matrice d'icônes. Un cas intéressant est également celui décrit dans (Tweedie et al. 1994)) où les coordonnées parallèles sont utilisées : les dimensions choisies sont par exemple la pertinence du document, sa date mais aussi les proportions de différents mots-clés. Les possibilités de sélection permettent alors de formuler très facilement des requêtes sur cet ensemble de documents.

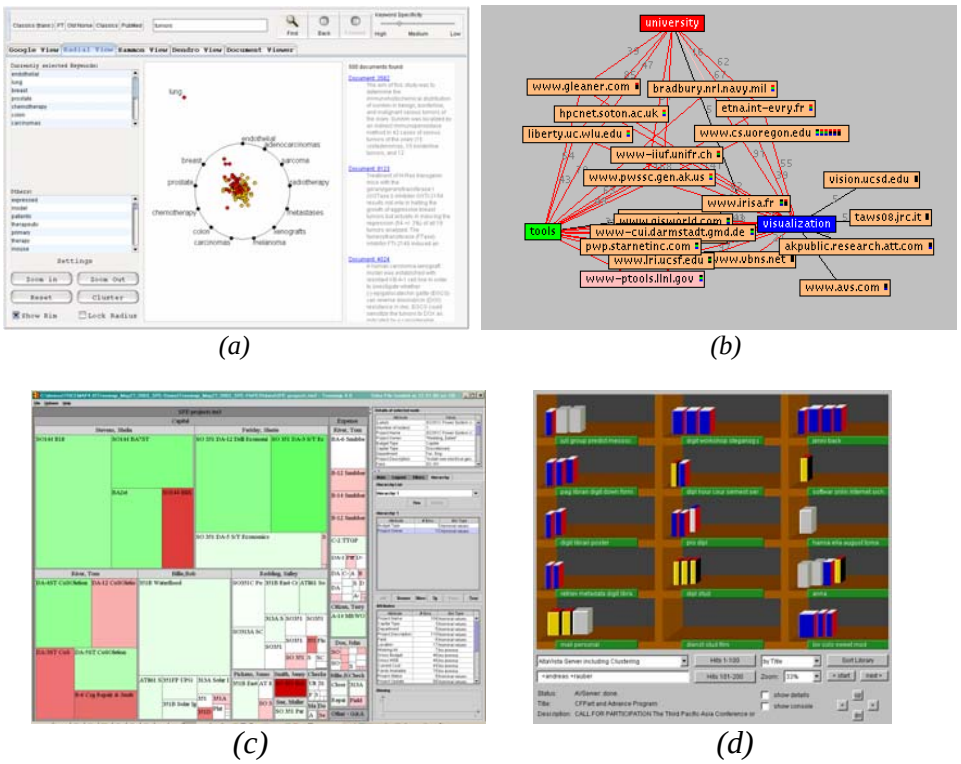
3. Choix des attributs visuels : le cas des métaphores

Nous avons vu dans les sections précédentes qu'il existe de multiples manières de représenter des données sous la forme de signes visuels. Cette représentation et les choix de codage dépendent de la tâche à accomplir. Cependant, suivant les choix qui sont effectués, un apprentissage plus ou moins long est nécessaire de la part de l'utilisateur afin de pouvoir utiliser correctement la visualisation. A titre d'exemple, on peut citer le cas des coordonnées parallèles qui permettent de détecter des corrélations linéaires entre deux attributs successifs sous la forme d'une certaine configuration visuelle, mais cependant la détection de cette configuration nécessite une forme d'apprentissage de la part de l'utilisateur. Une manière de minimiser ce temps d'apprentissage consiste à utiliser des représentations visuelles qui soient familières à l'utilisateur. Dans ce cadre, l'utilisation de métaphores visuelles représente un atout et un enjeu importants pour les méthodes de fouille visuelle. Le principe général consiste à choisir un environnement de visualisation (et de manipulation) qui soit une image d'un environnement réel connu de l'utilisateur. Ainsi, le système *PSDoom* utilise la métaphore d'un jeu vidéo dans lequel des monstres représentent des processus sous Unix que l'utilisateur peut « tuer » à l'aide de son arme. Nous détaillons donc maintenant quelques exemples de systèmes utilisant les métaphores.

ThemeScape est un exemple remarquable d'utilisation de la métaphore d'une île (Wise 1999) : au départ, il utilise une carte de documents qui ont été regroupés selon leur ressemblance thématique. Ensuite, il considère que les bords de la carte sont au « niveau de la mer » et donc représentés en bleu. L'élévation sur la carte est perçue par la couleur (du bleu, puis vert, puis « marron » et enfin blanc comme de la neige). Cette hauteur dépend de la pertinence des documents ou encore du nombre de documents présents dans un certain voisinage. Cette représentation est très intuitive pour l'utilisateur familier avec des cartes géographiques.

Le *« Perspective wall »* utilise l'image d'un mur en pseudo-3D (Mackinlay et al. 1991) : il s'agit d'un mur à 3 pans dont la partie centrale représente le point de focalisation et les deux pans extérieurs, dessinés en perspective, représentent le reste de la visualisation avec beaucoup moins de détails qu'au centre. Il est utilisé notamment pour représenter des documents.

LibViewer est aussi un bel exemple de réalisation (Rauber et Bina 2000) (voir figure 2) : à partir d'une classification de documents obtenue par une carte de Kohonen, ce système construit une bibliothèque virtuelle dans laquelle les étagères correspondent aux cellules de la carte et les documents traités sont représentés par des livres (avec différentes représentations visuelles suivant les types de documents). L'utilisateur a donc ainsi l'impression de manipuler une véritable bibliothèque.



(a) *(b)*

(c) *(d)*

Fig. 2 En (a) Visualisation dans le système Radial (Au et al., 2000). En (b) le système SQWID (McCrickard et Kehoe 1997) (images fournies par S. Rueger et S. McCrickard respectivement). En (c) un exemple d'une visualisation hiérarchique d'un ensemble de documents avec Treemap (Shneiderman et al. 2000) (image fournie gracieusement par le « University of Maryland Human-Computer Interaction Lab »). En (d) LibViewer du système SOMLib (Rauber et Bina 1999) (image fournie gracieusement par A. Rauber)

Enfin, un dernier exemple que nous avons choisi est le système « *City of News* » (Sparacino et al. 1999). Il permet de représenter des pages Web sous la forme d'une ville : les façades des gratte-ciels représentent les pages. Ces bâtiments peuvent ensuite être regroupés par quartier. L'utilisateur a alors l'impression de se promener dans une « ville » de documents qu'il aura bien souvent organisé lui-même (il s'agit par exemple de mémoriser les pages favorites de l'utilisateur).

4. Un domaine d'application : « Web usage mining »

Nous donnons ici un exemple de point de vue inverse de ceux choisis précédemment : nous considérons une application particulière, l'analyse des consultations d'un site Web, et nous montrons quelques méthodes pour visualiser les actions des internautes sur un site. Cette application est intéressante car elle met en oeuvre de nombreux types de données : en effet, lors de la consultation d'un site Web, on engendre des données numériques (e.g. temps passé sur une page), symboliques (e.g. pays d'origine) mais aussi des séquences (e.g. suite de pages), des dates, et enfin, on peut même considérer le contenu textuel des pages vues. Ainsi,

ce problème est intéressant pour les méthodes de fouille de données en général, ainsi que pour les méthodes visuelles car les données d'entrée sont très hétérogènes (et peuvent aussi être en volume important). Cette problématique concerne par ailleurs non seulement les sites mais plus généralement la « navigation » des utilisateurs dans des structures telles que des menus ou autres interfaces.

Webviz est certainement l'un des premiers systèmes dans ce domaine (Pitkow et Bhara 1994). Il représente le site par un graphe (les nœuds sont les pages, les arcs entre les nœuds représentent les transitions entre pages). Les nœuds sont répartis sur l'écran en fonction de leur profondeur dans le site. Ensuite, l'épaisseur des arcs entre les nœuds symbolise le nombre de passages. D'autres informations en couleur sont ajoutées notamment.

VisVIP (Cugini et Scholtz 1999) est un exemple de système plus récent qui visualise des informations supplémentaires (comme le temps passé sur chaque page). Il visualise l'ensemble du site sur un plan. Ensuite, il affiche avec une ligne courbe le chemin suivi par un internaute. L'information de durée de consultation d'une page est représentée par une barre verticale au-dessus de chaque nœud. Ce système propose une meilleure visualisation que la précédente cependant il ne peut pas afficher beaucoup de données car il est dépourvu de capacités de « généralisation ».

Chi et ses collègues (voir (Chi 2002)) ont une visualisation à base d'arbres pour représenter l'évolution des consultations d'un site Web. Les pages sont structurées sous la forme d'un arbre et l'épaisseur des liens correspond aux passages les plus importants. Sur la même visualisation il est possible de représenter également les résultats d'un algorithme de prédiction et de les comparer à la navigation réelle des utilisateurs.

L'approche que nous présentons maintenant se base sur une carte de Kohonen pour regrouper les pages (Benabdeslem et Bennani 2006) : ces regroupements ont lieu en fonction de la présence des pages dans les mêmes traces de navigation. La carte visualise donc des groupes et permet de suivre le chemin suivi entre ces groupes. Il en résulte que l'information est mieux « condensée » que dans les systèmes précédents, ce qui permet à ce système de traiter des données beaucoup plus volumineuses.

Enfin, nous pouvons mentionner d'autres approches utilisant des visualisations différentes, et notamment (Fabrikant et Buttenfield 1997) où l'on représente des groupes de « documents » vus simultanément sous la forme d'une carte 3D prenant la forme d'un paysage. Les documents proches les uns des autres ont été vus ensemble. La hauteur de la carte donne la fréquentation des documents.

5. Conclusions et perspectives

L'état de l'art présenté dans cet article est loin d'être exhaustif : nous avons choisi seulement quelques points de vue parmi ceux énumérés dans l'introduction, et même parmi ces points de vue, nous nous sommes limités à des méthodes bien établies. Nous espérons cependant que cela donne au lecteur une idée de la richesse des travaux dans ce domaine.

Parmi les autres points de vue qu'il nous semblerait important de traiter nous pouvons mentionner tout d'abord celui des types de données visualisées. En effet, nous n'avons pas décrit l'ensemble des types possible comme les séquences et autres données temporelles, mais aussi les graphes ou les matrices. De même nous n'avons pas abordé en détail la découverte interactive de connaissances (dans laquelle l'utilisateur guide un système d'apprentissage en fonction des visualisations intermédiaires qu'il obtient). Egalement le point de vue de la validation des méthodes est important et fait appel à un élargissement vers les sciences cognitives et psychologiques, puisque cela met en scène de vrais utilisateurs.

Enfin, les différentes méthodes d'interaction avec les visualisations mériteraient aussi d'être abordées.

Pour finir on peut relever que la fouille visuelle de données est un domaine à part entière qui s'étend dans de nombreuses directions laissant la place à une multitude de travaux. Les avancées réalisées depuis les premiers systèmes permettent d'envisager des perspectives très intéressantes, comme le traitement de grands volumes de données (jusqu'ici environ 1 million de données) ou encore l'utilisation d'interactions plus élaborées avec notamment la réalité virtuelle.

Références

(Azzag et al. 2006) Azzag H., F. Picarougne, C. Guinot et G. Venturini. VRMiner: a tool for multimedia databases mining with virtual reality. *Processing and Managing Complex Data for Decision Support* (2006). J. Darmont and O. Boussaid, Editors, ??? pages???

(Ankerst et al. 1999) Ankerst M., Christian Elsen, Martin Ester et Hans-Peter Kriegel, Visual Classification: An Interactive Approach to Decision Tree Construction, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 392-396, 1999

(Au et al. 2000) Au, P., M. Carey, S. Sewraz, Y. Guo, et S. M. Rüger (2000). New paradigms in information visualization. In Research and Development in Information Retrieval, pp. 307–309.

(Becker et Cleveland 1987) Becker, R. A. et W. S. Cleveland (1987). Brushing Scatterplots. Technometrics 29, 127–142. Reprinted in Dynamic Graphics for Data Analysis, edited by W. S. Cleveland and M. E. McGill, Chapman and Hall, New York, 1988.

(Benabdeslem et Bennani 2006) Benabdeslem K. et Bennani Y., Classification et visualisation des données d'usages d'Internet, Atelier FW-EGC'06, Fouille du Web, Conférence EGC'2006. ???pages???

(Blake et Merz 1998) C.L. Blake et C.J. Merz (1998). UCI Repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science. 1998.

(Brunk et al. 1997) Brunk C., Kelly J. et Kohavi R. (1997), MineSet: an integrated system for data mining, International Conference on Knowledge Discovery and Data Mining (KDD'97), AAAI Press, pp 135-138.

(Da Costa et Venturini 2006) Da Costa D. et G. Venturini (2006). An Interactive Visualization Environment for Data Exploration Using Points of Interest. (ADMA2006: the 2nd International Conference on Advanced Data Mining and Applications), August 14-16 2006, Xi'An, China.

(Chernoff 1973) Chernoff H., The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68 : 361-368, 1973.

(Chi et Riedl 1998) Chi E. et Riedl J., An Operator Interaction Framework for Visualization Systems. Actes de la conférence InfoVis '98 pp. 63-70.

(Chi 2002) Chi E., Improving Web Usability Through Visualization, *IEEE Internet Computing*, 2002, pp. 64-71.

(Cugini et Scholtz 1999) Cugini J. et Scholtz J., VISVIP: 3D Visualization of Paths through Web Sites, WebVis'99.???pages et details???

(Eick et al. 1992) Eick, S. G., Steffen J. L., Sumner, Eric E.: SeeSoft: A tool for visualizing line oriented-software statistics In: IEEE Transactions on Software Engineering, 18 (1992) 11, p. 957-968.

(Fabrikant et Buttenfield 1997) Fabrikant, S. I. et Buttenfield, B. P., Envisioning User Access to a Large Data Archive. Proceedings, GIS/LIS '97, 686-692.

(Fekete et Plaisant 2002) Fekete Jean-Daniel et Catherine Plaisant, Interactive Information Visualization of a Million Items, Proceedings of IEEE Symposium on Information Visualization 2002 (InfoVis 2002), IEEE Press, pages 117-124, Boston, USA, Octobre 2002.

(Fisher 1936) Fisher R.A., The Use of Multiple Measurements in Taxonomic Problems, *Annals of Eugenics*, 7, 1936, 179 -188.

(Fua et al. 1999) Fua Ying-Huey, Matthew O. Ward et Elke A. Rundensteiner, Hierarchical Parallel Coordinates for Exploration of Large Datasets, http://www.dgp.toronto.edu/~ravin/courses/csc2524/fua_vis99.pdf

(Green et al. 1989) Green, P. E., Carmone, F. J., et Smith, S. M. (1989). Multidimensional scaling: concepts and applications. Boston: Allyn & Bacon.

(Havre et al. 1999) Havre Susan, Hetzler Beth, et Nowell Lucy. ThemeRiver : In Search of Trends,Patterns, and Relationships. Présenté au IEEE Symposium on Information Visualization, InfoVis'99, http://www.pnl.gov/infoviz/themeriver99.pdf

(Hearst 1995) Hearst, Marti A.: TileBars: Visualization of Term Distribution Information in Full Text Information Access. In: Katz, Irvin R.; Mack, Robert L.; Marks, Linn et al. (Eds.): CHI 1995: Conference Proceedings Human Factors in Computing Systems. Conference: Denver, CO, May 7-11 1995. New York (ACM Press) 1995. p. 59-66.

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

(Inxight 1999) Hyperbolic Trees, Inxight: http://www.inxight.com. 1999.

(Janey 1992) Janey Nicolas, Modélisation et synthèse d'images d'arbres et de bassins fluviaux associant méthodes combinatoires et plongement automatique d'arbres et cartes planaires, Thèse de Doctorat, Université de Franche-Comté, 1992.

(Kohonen 1989) Kohonen T., Self-Organization and Associative Memory. Springer-Verlag, Berlin, 3rd edition, 1989.

(Kohonen 1998) Kohonen, T, Self-organization of very large document collections: State of the art. In: Niklasson, Lars; Bodén, Mikael; Ziemke, Tom (Eds.): Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks. Conference: Skövde, Sweden, 1998 September 2-4 Springer (London) 1998. p. 65-74, vol. 1.

(Korfhage 1991) Korfhage, R., To see, or not to see: Is that the query? In A. Bookstein, Y. Chiaramella, G. Salton, et V. V. Raghavan (Eds.), Proceedings of the 14th Annual International ACM SIGIR, Conference on Research and Development in Information Retrieval. Chicago, Illinois, USA, October 13-16, 1991 (Special Issue of the SIGIR Forum), pp. 134–141. ACM. http://mmis.doc.ic.ac.uk/www-pub/npiv-sigir2000.pdf

(Leung et Apperley 1994), Leung Y. K. et Apperley M. D., " A Review and Taxonomy of Distortion-Oriented Presentation Techniques". Journal ACM Transaction on Computer-Human Interaction. 1(2), Juin 1994, pp. 126-160.

(Mackinlay et al. 1991) Mackinlay, Jock D.; Robertson, George G.; Card, Stuart K.: The Perspective Wall: Detail and Context Smoothly Integrated. In: Robertson, S. P.; Olson, G. M.; Olson, J.S. (Eds.): CHI 1991: Conference Proceedings Human Factors in

Computing Systems. Conference: New Orleans, LA, April 27 - May 2 1991. New York (ACM Press) 1991.p. 173-179.

(McCrickard et Kehoe 1997) McCrickard, D. Scott; Kehoe, Colleen M.: Visualizing Search Results using SQWID. In: WWW 6: Sixth International World Wide Web Conference. Conference: Santa Clara, CA, April 7 - 11 1997.

(Mokaddem et al. 2006) Mokaddem F., F. Picarougne, H. Azzag, C. Guinot, G. Venturini (2006). Techniques visuelles de recherche d'informations sur le Web. Revue des Nouvelles Technologies de l'Information (RNTI), numéro spécial Visualisation en Extraction des Connaissances, Pascale Kuntz et François Poulet rédacteurs invités, Cépaduès édition, pages 21-47.

(Munzner et Burchard, 1995) Munzner Tamara et Paul Burchard. Visualizing the structure of the world wide web in 3d hyperbolic space. In Proceedings of the first symposium on Virtual reality modeling language, pages 33–38. ACM Press, 1995.

(Nowell et al. 1996) Nowell, L.; France R.; Hix D., Health L., Fox E., Visualizing Search Results: Some Alternatives to Query-Document Similarity. In: Frei, Hans-Peter; Harman, Donna K.; Schäuble, Peter et al. (Eds.): SIGIR 1996: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Conference: Zürich, Switzerland, August 18 -22 1996. New York (ACM Press) 1996. p. 67-75.

(Ogden et al., 1998) William C. Ogden, Mark W. Davis, et Sean Rice. Document thumbnail visualization for rapid relevance judgments: When do they pay off? In Ellen M. Voorhees et Donna K. Harman, editors, Proceedings of TREC-7, 7th Text Retrieval Conference. National Institute of Standards and Technology, Gaithersburg, US, nov 1998.

(Pickett et Grinstein 1988) Pickett Ronald M. et Georges Grinstein. Iconographic displays for visualizing multidimensional data. In Proceedings of the 1988 IEEE International Conference on Systems, Man, and Cybernetics, volume 1, pages 514-519, 1988.

(Pitkow et Bhara 1994) Pitkow J. et K. Bharat. Webviz: A tool for worldwide web access log visualization. Proceedings of the First International World-Wide Web Conference, Geneva, Switzerland, May 1994.

(Rauber et Bina 2000) Rauber A., Bina H. (2000), An Old-Fashioned Approach to Web Search Results Visualization. In: Tjoa, A Min; Wagner, Roland R.; Al-Zobaidie, Ala (Eds.): Proceedings 11th International Workshop on Database and Expert Systems Applications. 2000. p. 615-619.

(Robertson et al. 1991) Robertson George, Jock Mackinlay, and Stuart Card. Animated 3d visualizations of hierarchical information. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems, pages 189-194. ACM, April 1991.

(Shneiderman et al. 2000) Shneiderman B., Feldman D., Rose A., et Ferre Grau X., Visualizing digital library search results with categorical and hierarchical axes. In ACM Digital Libraries, 2000.

(Sparacino et al. 1999) Sparacino, F., R. DeVaul, C. Wren, G. MacNeil, G. Daveport, and A. Pentland. City of News, in SIGGRAPH 99, Visual Proceedings, Emerging Technologies. 1999.

(Tesler et Strasnick 1992) Tesler Joel et Steve Strasnick. Fsn: The 3d file system navigator.

(Thearling et al. 1998) Thearling K, Becker B., DeCoste D., Mawby B., Pilote M. et Sommerfield D. (1998), Visualizing data mining models, Proceedings of the Integration of Data Mining and Data Visualization Workshop, Springer Verlag, 1998.

(Tweedie et al. 1994) Tweedie, L. A.; Spence, R.; Williams, D. Bhogal R.: The Attribute Explorer. Video Track. In: Adelson, B.; Dumais, S.; Olson, J. S. (Eds.): CHI 1994: Conference Proceedings Human Factors in Computing Systems. Conference: Boston, MA, April 24-28 1994. New York (ACM Press) 1994. p. 435-436.

(Wise 1999) Wise, J. A.: The Ecological Approach to Text Visualization. In: Journal of the American Society for Information Science (JASIS), 50 (1999) 13, p. 1224-1233.

(Wong et Bergeron 1997) Wong, P. C., et Bergeron, R. D., 30 years of multidimensional multivariate visualization. Scientific Visualization Overview, Methodologies, Techniques. IEEE Computer Society Press.

summary

We present in this paper an overview of visual data mining (VDM). First we motivate this approach with respect to other methods and we remind the reader with its foundations. We show that there are numerous possible view points on VDM: the data or knowledge to be visualized, the task to be fulfiller, the chosen visual representation, the computation method or the application domain. We have selected first the data point of view and we detail representative approaches for visualizing numeric, hierarchical or textual data. Then we consider the visual representation point of view and describe typical examples of metaphors used in VDM. Finally, we consider a specific application domain, Web usage mining, and then we conclude on the possible perspectives of VDM.