Revue des Nouvelles Technologies de l'Information
Sous la direction de Djamel A. Zighed et Gilles Venturini

RNTI-E-7

Visualisation en extraction des connaissances

Rédacteurs invités :

François Poulet
(ESIEA - Pôle ECD, Laval)

Pascale Kuntz
(LINA, Nantes)

CÉPADUÈS-ÉDITIONS
111, rue Vauquelin
31100 TOULOUSE – France
Tél. : 05 61 40 57 36 – Fax : 05 61 41 79 89
(de l'étranger) + 33 5 61 40 57 36 – Fax : + 33 5 61 41 79 89
www.cepadues.com
courriel : cepadues@cepadues.com

Chez le même éditeur

RNTI-Revue des Nouvelles Technologies de l'Information
Sous la direction de Djamel A. Zighed et Gilles Venturini

n°1 : Entreposage fouille de données
E1 : Mesures de qualité pour la fouille de données
E2 : Extraction et gestion des connaissances EGC 2004
C1 : Classification et fouille de données
E3 : Extraction et gestion des connaissances EGC 2005
B1 : 1re Journée Francophone sur les Entrepôts de Données et l'Analyse en ligne EDA 2005
E4 : Fouille de données complexes
E5 : Extraction des connaissances : Etat et perspectives
E6 : Extraction et gestion des connaissances EGC 2006

ISBN : 2.85428.733.9

Dépôt légal : mars 2006

N° éditeur : 733

LE MOT DES DIRECTEURS DE LA COLLECTION RNTI

Chères Lectrices, Chers Lecteurs,

La Revue des Nouvelles Technologies de l'Information a pour objectif d'être un outil de communication et d'échange scientifique dans le monde francophone. Nous avons voulu rendre cette publication accessible au plus grand nombre de chercheurs francophones universitaires ou non . Le succès d'une telle initiative repose essentiellement sur la qualité des contributions. Pour nous en assurer, nous avons imposé aux rédacteurs invités une charte et des exigences fortes.

Pour répondre aux besoins des lecteurs qui cherchent des articles souvent pointus dans leur domaine de spécialité, nous avons structuré la publication selon différents thèmes (listés ci-dessous). RNTI accueille deux types de numéros :

- des actes de conférences sélectives garantissant une haute qualité des articles (par exemple, nous demandons à ce que trois relecteurs émettent un avis sur les articles soumis). Ainsi le numéro RNTI-B-1 a concerné les actes de EDA'2005 (Entrepôts de Données et l'Analyse en ligne), et nous venons de publier les actes de la conférence EGC'2006 (Extraction et Gestion des Connaissances).
- des numéros à thème faisant l'objet d'un appel à communication. Chaque numéro à thème est édité par un ou plusieurs rédacteurs en chef invités. Un comité de lecture d'une quinzaine de personne est formé à cette occasion pour évaluer les papiers avec une exigence de trois avis par papier soumis. Est paru tout récemment un numéro spécial sur la fouille de données complexes. Nous avons aujourd'hui le plaisir d'accueillir ce numéro sur la Visualisation en Extraction des connaissances édité par Pascale Kuntz et François Poulet. Cela correspond à un thème suscitant beaucoup de perspectives et aussi d'attente de la part de la communauté scientifique.

Les thèmes de RNTI sont ainsi classés de la manière suivante :

- RNTI - A : Apprentissage
- RNTI - B : Bases de données
- RNTI - C : Classification
- RNTI - E : Extraction et Gestion des Connaissances
- RNTI - S : Statistiques
- RNTI - W : Web

Nous espérons vivement que ce numéro vous donnera à toutes et à tous une entière satisfaction. Pour tout renseignement, nous vous invitons à consulter notre site Web et à nous contacter. En particulier, nous sommes à votre écoute pour toute proposition de nouveaux numéros spéciaux.

Djamel A. Zighed et Gilles Venturini.
http ://www.antsearch.univ-tours.fr/rnti

PRÉFACE

1 La Visualisation en ECD

L'essor en cette dernière décennie des recherches sur la Visualisation en Extraction de Connaissances dans les Données (ECD) s'ancre dans une longue tradition ponctuée d'évènements significatifs ; sans remonter aux premiers travaux du XVIII^{eme} siècle rappelés par Tufte [1983], citons pour la période contemporaine, le congrès sur l'information visuelle de Milan de 1961 qui projetait de créer un conseil international de la recherche sur l'information visuelle (Palsky and Robic [2000]), la Graphique de Bertin [1967], l'Exploratory Data Analysis de Tukey [1977], et plus récemment par exemple les ouvrages de Chambers et al. [1983] ou Cleveland [1985].

Si les recherches des années 60 et 70 ont permis d'établir des bases conceptuelles, les avancées récentes en psychologie cognitive sur le traitement visuel de l'information (e.g. Ware [2000]), et en informatique sur l'accessibilité de plus en plus aisée à des outils de restitution de grande qualité, conduisent aujourd'hui à un regain d'intérêt majeur pour la représentation visuelle des données et des connaissances.

Bien qu'intimement liée par essence aux travaux à plus large spectre en Visualisation de l'information (Keim [2002]), la Visualisation en ECD cherche à circonscrire son champ spécifique. Cela peut se faire partiellement en déclinant les besoins associés aux différentes étapes d'un processus d'ECD.

Dans la démarche la plus courante, la visualisation n'intervient de façon majeure qu'à deux étapes du processus : en amont, pour représenter les données, et en aval pour restituer les connaissances extraites.

Dans la phase amont, les données considérées généralement en entrée sont des objets décrits par des variables de différente nature qui peuvent provenir de bases de données relationnelles. La problématique rentre alors dans le cadre du modèle proposé par Card et al. [1999] qui se base en partie sur les travaux précédemment cités. Dans ce modèle, les transformations graphiques s'appuient sur une grammaire qui associe à chaque variable décrivant les données une variable visuelle (forme, couleur, position, ...). Les transformations sur la vue sont associées aux opérateurs de manipulation accessibles à l'utilisateur (translations, rotations, overview + details, focus + context, ...). Les approches proposées sont soit génériques et visent des types de variables variés, ou bien spécifiques en se spécialisant sur des données particulières (Chi [2000]). Les représentations visuelles permettent ici d'offrir une vue synthétique des données, d'émettre des hypothèses (analyse exploratoire, Keim and Kriegel [1996]), ou de valider visuellement des hypothèses pré-établies (analyse confirmatoire).

Dans la phase en aval, il s'agit de représenter visuellement les structures découvertes par un algorithme -souvent automatique- de fouille. Le terme « structure » renvoie à des concepts très variés et relatifs aux approches utilisées : arbres de décision, partitions, réseaux de règles

d'association, ... Les méthodes de visualisation sont alors intrinsèquement liées à la structure à représenter.

Au-delà des objectifs initiaux visés pour ces deux phases, la Fouille visuelle de données qui, en tant que thème à part entière a émergé dans les années 90, consiste à coupler plus étroitement la visualisation avec les processus analytiques afin de développer des nouveaux outils qui bénéficient des avantages de chacun (Wong [1999]). De cet objectif découle un besoin croissant d'interactivité avec les utilisateurs. Dès lors, ces derniers ne sont plus considérés comme des entités externes qui règlent *a priori* les paramètres et analysent *in fine* les résultats obtenus ; ils sont intégrés comme une composante à part entière du processus de fouille (Do and Poulet [2003]). Cette approche coopérative centrée utilisateur permet de :

- s'appuyer sur les connaissances internes de l'expert des données lors du processus de fouille ;
- d'augmenter, de par sa participation active dans son élaboration, la confiance de l'utilisateur dans les modèles proposés lors du processus ;
- de développer de nouvelles heuristiques performantes qui exploitent les capacités cognitives en reconnaissance des formes.

Ainsi, le duo Visualisation + Interaction est certainement un concept majeur sur lequel devra s'appuyer l'ECD pour affronter les deux challenges majeurs actuels : la fouille de données complexes où les données proviennent de sources variées (textes, mutimedia, ...) et ne se représentent plus aisément sous la forme de tableaux Objets $\times$ Variables, et le passage à l'échelle inhérent aux capacités de stockage des données (doublement tous les 9 mois).

2 Contenu de l'ouvrage

Du fait de l'histoire récente de la Visualisation en ECD les références du domaine apparaissent plutôt dans les publications d'ateliers spécifiques (*e.g.* International Workshop on Visual Data Mining ((ECML/PKDD'01, 02, ICDM'03)) ou de revues en visualisation (*e.g.* IEEE Trans. on Visualization and Computer Graphics, IEEE Computer Graphics and Applications, Information Visualization) et les ouvrages spécifiques sont encore très rares (Fayyad et al. 02).

Cet ouvrage, restreint à la communauté francophone, fait suite à deux ateliers organisés lors de la conférence annuelle EGC (Extraction et Gestion des Connaissances). La présence dès le premier atelier à Lyon en 2002 de près d'une quarantaine de participants était le signe de l'intérêt porté à cette problématique dans la communauté scientifique. Les articles de cet ouvrage présentent les travaux menés dans une douzaine de laboratoires français[1], certains en relation étroite avec des entreprises, et leurs thématiques balayent les différentes approches abordées dans notre introduction. Plus précisément, les contributions peuvent être organisées en deux grandes parties.

[1]Centre de Recherche Economique de l'Université de Saint-Etienne - ESIEA Pôle ECD, Laval - Equipe de Recherche Economique de l'Université de Saint-Etienne - Groupe de Recherche en Economie Mathématique et Quantitative, Toulouse - Institut de Recherche en Informatique de Toulouse - Institut de Recherche en Informatique et Systèmes Aléatoires, Rennes - Laboratoire d'Informatique de Paris 6 - Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes, Clermont Ferrand - Laboratoire d'Informatique de Nantes Atlantique - Laboratoire d'Informatique en Image et Systèmes d'Information, Lyon - Laboratoire d'Informatique de l'Université de Tours - Laboratoire d'Etudes et de Recherche sur l'Economie, les Politiques et les Systèmes Sociaux, Toulouse.

La première partie, généraliste, présente à travers des états de l'art à la fois les propriétés mais aussi les limites des outils de visualisation actuels et suggèrent de nouvelles voies de recherche. B. Legrand et M. Soto cherchent, après une analyse critique de l'existant, à définir les bases d'un outil générique resistant au facteur d'échelle, c'est à dire capable de gérer des données très volumineuses tout en limitant les effets induits par les hypothèses implicites associées aux pré-traitements. F. Mokaddem *et al.* présentent un état de l'art sur les techniques visuelles utilisées en recherche d'information documentaire. Ce dernier intègre les aspects cognitifs et méthodologiques qui peuvent s'étendre à d'autres problématiques, et évalue également les aspects logiciels.

La seconde partie est orientée vers les applications : classification, recherche de règles, analyse de séquences, détection d'objets atypiques, et intégration des données spatiales.

- La classification et l'analyse des données ont depuis longtemps adjoint aux structures qu'elles étudient des représentations visuelles (arbres hiérarchiques, pyramides, cartes factorielles, ...). T. Do et F. Poulet proposent ici une nouvelle approche de construction interactive et d'interprétation graphique de séparateurs à vaste marge (SVM) en vue d'applications sur des grands ensembles de données. R. Priam développe une extension des cartes auto-organisatrices de Kohonen pour des vecteurs textuels qui débouche sur une méthode de visualisation synthétique d'un corpus de documents textuels.

- La très grande majorité des systèmes opérationnels d'ECD portant sur des enregistrements de bases de données relationnelles, la recherche de dépendances et de relations d'implication ont été certainement l'un des domaines les plus féconds de la littérature de l'ECD de cette dernière décennie. Pour améliorer la compréhension des dépendances fonctionnelles et d'inclusion extraites par des algorithmes automatiques, F. de Marchi et J.-M. Petit proposent de fournir en complément à l'utilisateur un ensemble d'exemples issus de la base de données qui illustrent ces relations. Pour contourner les difficultés d'interprétation inhérentes aux algorithmes automatiques de fouille de règles d'association, P. Kuntz *et al.* développent une approche interactive intégrant un modèle graphique qui permet à l'utilisateur de piloter le processus de fouille en jouant le rôle d'une heuristique locale.

- De nombreuses problématiques applicatives (analyse de logs, séquences moléculaires, traitement de la langue, ...) font intervenir comme modèle sous-jacent des séquences discrètes. C. Largeron et C. Dreissia présentent un outil de visualisation de séquences sous forme d'arbres de suffixes probabilistes (PST) qui permet de comparer les séquences entre elles et de faciliter les classements des nouvelles dans un cadre supervisé.

- L'augmentation des dimensions manipulées en ECD renouvelle le problème de la détection des éléments "atypiques" pour des dimensions caractéristiques. L. Boudjeloud et F. Poulet proposent une méthode graphique, basée sur un algorithme génétique, pour d'une part déceler les éléments atypiques et d'autre part tenter d'expliquer la nature (bruit, différence significative, ...) de sa singularité.

- L'analyse de données hétérogènes, car souvent issues de sources multiples, faisant intervenir simultanément des dimensions spatiales et temporelles est une problématique

encore peu explorée en ECD. S. Coehlo *et al.* décrivent le résultat d'un travail interdisciplinaire qui vise à mieux appréhender, dans une zone géographique donnée, l'influence réciproque entre la structuration d'un territoire et les comportements des acteurs. L'accent est mis ici sur la description d'une plate-forme de simulation et de visualisation qui permet la modélisation et l'analyse des comportements.

Pascale Kuntz
François Poulet

Références

J. Bertin. *La graphique et le traitement graphique de l'information*. Flammarion, 1967. Réédition de l'Ecole des Hautes Etudes en Sciences Sociales.

S. Card, J. MacKinlay, and B. Shneiderman. *Readings in Information Visualization : Using Vision to Think*. Morgan Kaufman Pub., 1999.

J. Chambers, W.S. Cleveland, B. Kleiner, and P. Tukey. *Graphical methods for data analysis*. Wadsworth, 1983.

E. Chi. A taxonomy of visualization techniques using the data state reference model. In *Proc. of IEEE Symp. on Information Visualization (Infoviz)*, pages 69–76, 2000.

W.S. Cleveland. *The elements of graphing data*. Wadsworth, 1985.

T.N. Do and F. Poulet. Interactive visualization tools for visual data mining. In *Proc. of Human Centered Processes Conf.* 14^{th} *mini-EURO Conf.*, pages 299–304, 2003.

D.A. Keim. Information visualization and visual data mining. *IEEE Trans. on Visualization and Computer Graphics*, 8(1) :1–8, 2002.

D.A. Keim and H.P. Kriegel. Visualization techniques for mining large databases : a comparison. *IEEE Trans. on Knowledge and Data Engineering*, 8(6) :923–938, 1996.

G. Palsky and M.C. Robic. Aux sources de la sémiologie graphique, colloque 30 ans de sémiologie graphique. *Cybergeo*, (144), 2000.

E. Tufte. *The visual display of quantitative information*. Graphics Press, 1983.

J.W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

C. Ware. *Information visualization : Perception for design*. Interactive Technologies, Morgan Kaufmann, 2000.

P.C. Wong. Visual data mining. *IEEE Computer Graphics and Applications*, 19(5) :20–21, 1999.

Comité de lecture du numéro :

N. Belkhiter (Université de Laval, Québec)
H. Briand (LINA, Nantes)
C. Djeraba (LIFL, Lille)
B. Le Grand (LIP6, Paris)
G. Mélançon (LIRMM, Montpellier)
M. Noirhomme-Fraiture (Université de Namur, Belgique)
M. Soto (LIP6, Paris)
G. Venturini (LI, Tours)
D. Zighed (ERIC, Lyon-2)

Relecteurs additionnels :

L. Boudjeloud (ESIEA - Pôle ECD, Laval)
T.-N. Do (University of Can Tho, Vietnam)
R. Lehn (LINA, Nantes)

TABLE DES MATIÈRES

Visualisation exploratoire, généricité, exhaustivité et facteur d'échelle

Bénédicte Le Grand et Michel Soto

Laboratoire d'Informatique de Paris 6
8, rue du Capitaine Scott 75015 Paris
{Benedicte.Le-Grand ; Michel.Soto}@lip6.fr

Résumé. L'objectif de cet article est d'étudier les limites des outils de visualisation exploratoire actuels et d'explorer de nouvelles voies pour dépasser ces limites. Ce travail est mené selon deux axes fortement liés :

- *Impact du facteur d'échelle* : les outils actuels, procèdent par isolation, simplification et réduction, afin de ramener la masse de données vers un niveau acceptable. Ce mode opératoire a une dimension arbitraire qui fait disparaître des liens d'ordre structurel et sémantique, ce qui réduit les chances d'extraction d'information.
- *Impact du paradigme* : regard critique sur le choix en termes d'hypothèses et prétraitements sous-jacents -tel que le filtrage- effectués par ces outils sur les données elles-mêmes. En effet, ces hypothèses et prétraitements conditionnent fortement, par leur effet réducteur, les informations que les outils pourront ou non extraire des données.

Les deux points précédents posent la question de la *généricité*, de l'*exhaustivité* et de la *neutralité* des outils de visualisation. La prise en compte de ces contraintes nous fournit les bases pour des outils de visualisation exploratoire génériques, exhaustifs et pouvant résister au facteur d'échelle.

1. Introduction

Les systèmes d'information actuels regorgent de données de tout type et la difficulté réside maintenant dans l'exploitation et l'interprétation de ces données pour en extraire des informations, puis des connaissances. De nombreux outils de fouille de données ont été développés pour atteindre cet objectif ; dans le cadre de nos recherches, nous nous intéressons tout particulièrement aux outils de *visualisation exploratoire des données*, qui permettent d'utiliser de manière intuitive les facultés cognitives humaines lors du processus de visualisation [Keim, 2002]. L'*analyse exploratoire*, contrairement aux techniques de présentation ou de confirmation [Ankerst, 2000], traite des données sur lesquelles l'utilisateur n'a fait aucune hypothèse. Grâce à une exploration interactive, l'utilisateur forme des hypothèses, ensuite révélées par une visualisation appropriée. Ce mode de visualisation doit permettre de déceler parmi les données des informations implicites mais potentiellement utiles, en particulier les relations entre les données, qui sont essentielles à l'élaboration des connaissances.

L'objectif de cet article est d'étudier les limites des outils de visualisation actuels lorsque le volume des données augmente et de spécifier les bases d'un outil générique *résistant au*

Techniques visuelles de recherche d'information

Fewzi Mokaddem*, Fabien Picarougne*, Hanene Azzag*,
Christiane Guinot* **, Gilles Venturini*

*Laboratoire d'Informatique de l'Université de Tours,
École Polytechnique de l'Université de Tours - Département Informatique,
64, Avenue Jean Portalis, F-37200 Tours.
fewzi.mokaddem@etu.univ-tours.fr,
{fabien.picarougne, hanene.azzag, venturini}@univ-tours.fr,
**CE.R.I.E.S.,
20 rue Victor Noir, F-92521 Neuilly-sur-Seine Cedex
christiane.guinot@ceries-lab.com.

Résumé. Nous exposons dans cet article un état de l'art sur les techniques visuelles pouvant être utilisées dans la recherche d'information documentaire sur Internet ou dans un système d'information. Nous détaillons dans un premier temps les techniques qui permettent aux utilisateurs de formuler et d'affiner leurs requêtes de façon interactive et visuelle. Nous présentons ensuite les techniques permettant de représenter un document isolément puis les techniques et systèmes permettant de représenter un ensemble de documents. Nous analysons les atouts et faiblesses de ces méthodes et nous dégageons des perspectives pour ce domaine prometteur.

1. Introduction

La quantité d'information offerte au public sur le Web est très importante et augmente sans cesse. Il est certain que les moteurs de recherche à interface texte dite « classique » vont devenir de moins en moins efficaces face à cette inflation de l'information. En général, avec ce genre de méthodes classiques, les interfaces sont toutes composées sur le même modèle : un champ texte associé généralement à une ligne pour saisir la requête, un bouton pour lancer la recherche, un affichage des résultats sous forme de listes composées de centaines et parfois de milliers de documents relatifs aux critères. L'utilisateur vérifie les résultats et reformule la requête plusieurs fois pour trouver ce qu'il cherche. Une étude [Jansen et al., 1998][1] a montré qu'en moyenne 20 à 30 documents sont explorés par les utilisateurs, car ils se basent généralement sur le classement. Or ce n'est pas toujours un moyen efficace pour trouver le résultat souhaité rapidement.

Durant cette dernière décennie plusieurs travaux ont été menés en perception visuelle comme par exemple [Myers, 2000] qui a montré que l'être humain a une perception d'abord globale (gestalt-perception) d'une scène, avant de porter son attention aux détails. Ce type de travaux a guidé d'autres recherches sur l'analyse concurrentielle, le data mining, la cartographie de fonds documentaires, vers la conception de nouvelles techniques pour parer les faiblesses des anciennes méthodes, et particulièrement la recherche d'information sur le Web. Des systèmes ont émergé ces dernières années dans le but de permettre aux utilisateurs

[1] 86% de 18 113 utilisateurs n'ont consulté pas plus que les trois premières pages résultant d'Excite (avec 10 liens par page), 77% pas plus de deux pages et 58% pas plus d'une page.

Vis-SVM : approche coopérative en fouille de données

Thanh-Nghi Do, François Poulet

ESIEA – Pôle ECD
38, rue des Docteurs Calmette et Guérin
Parc Universitaire de Laval - Changé
53000 Laval
(dothanh, poulet)@esiea-ouest.fr

Résumé. La compréhension des résultats en sortie d'un algorithme de fouille de données est aussi importante que d'obtenir de bons taux de précision. Malheureusement, les modèles obtenus par les algorithmes de support vector machines ou séparateurs à vaste marge (SVM) fournissent seulement les vecteurs support qui sont utilisés comme une « boîte noire » pour classifier efficacement les données avec de bons taux de précision. Il est donc indispensable d'améliorer la compréhensibilité des modèles de SVM. Cet article présente différentes coopérations entre des méthodes de visualisation et des algorithmes de SVM en fouille de données. En post-traitement d'algorithmes de SVM, nous présentons une approche coopérative graphique interactive pour interpréter des résultats de classification, régression et détection d'individus atypiques. Nous étendons l'approche d'interprétation graphique pour améliorer les résultats obtenus par la classification de SVM. Nous présentons ensuite une approche coopérative permettant d'impliquer plus significativement l'utilisateur dans la tâche de classification à l'aide de SVM. Ce type d'approche présente notamment comme avantage la possibilité d'utiliser les capacités humaines en reconnaissance de formes par le biais de méthodes de visualisation. L'utilisateur a une meilleure compréhension du modèle construit et une meilleure confiance dans ce modèle parce qu'il a participé activement à sa construction. Nous montrons comment l'utilisateur peut utiliser des outils coopératifs pour construire des modèles de SVM. Une étape de prétraitement est également utilisée dans notre outil coopératif pour pouvoir traiter de grands ensembles de données. Nous évaluons les performances de la nouvelle approche coopérative sur les ensembles de données de l'UCI, Delve, Statlog et biomédicales.

1. Introduction

La fouille de données est un domaine récent de l'informatique dont le développement est lié aux masses de données de plus en plus importantes qui sont stockées à l'heure actuelle. 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 ». Ce processus est complexe, il vise à exploiter des techniques venant de différents domaines de recherches (intelligence artificielle, apprentissage automatique, statistique, analyse de données, visualisation d'informations, bases de données) pour l'extraction de connaissances. Parmi elles, on trouve les arbres de

CASOM : Carte auto-organisée pour l'analyse exploratoire des tableaux de contingence

Rodolphe Priam
IRISA - Projet TexMex
263 av Gén Leclerc F-35000 Rennes
rpriam@gmail.com

Résumé. La visualisation des connaissances par des méthodes d'extraction de l'information pour un corpus de données multimédia est une question pertinente aussi bien en recherche d'information où l'on cherche les meilleurs documents répondant à une requête, qu'en analyse des données où l'on cherche à comprendre quantitativement le contenu du corpus. En effet, en recherche d'information, les corrélations inter-variables permettent d'enrichir la requête de la même façon qu'elles renseignent sur les liaisons interprétables en analyse des données. Une manière générale de répondre à l'objectif posé est l'emploi de méthodes de réduction efficaces qui permettent de mettre en évidence les différentes caractéristiques principales et locales du corpus. Les méthodes de carte auto-organisatrice entrent dans cette optique tout en apportant la dimension supplémentaire d'une carte projective de la distribution et partitionnant le plan en diverses thématiques adjacentes. Ces méthodes rendent appréhendable par l'humain un nuage de points plongé dans un espace de grande dimension par la construction d'une surface discrète qui épouse la forme de sa distribution. Elles offrent ainsi une propriété de cartographie apte à montrer une structure sous-jacente. Dans ce cadre, nous développons une représentation originale des cartes de Kohonen pour des vecteurs textuels bruts. Nous fournissons des indicateurs numériques interprétables et aboutissons à la définition d'une méthode de visualisation synthétique et globale d'un corpus : un *biplot* sous la forme d'un graphe de mots révélant leurs liaisons statistiques, superposable à la projection des documents. La méthode est illustrée par le graphe du vocabulaire extrait d'un corpus de données réelles.

1 Introduction

Dans ce papier, nous introduisons les algorithmes de carte auto-organisatrice et décrivons les principales représentations de cartes auto-organisatrices présentes dans la littérature. Ce cadre posé, nous développons notre méthode CASOM adaptée aux matrices de comptage et mettons en évidence ses diverses propriétés particulières en terme de critère optimisé et métrique. La méthode est illustrée sur un corpus de résumés textuels en construisant des graphes qui montrent les liaisons statistiques entre les termes ou mots du vocabulaire sélectionné dans le corpus ; le graphe de mots a la propriété de se superposer à celui des documents. Cette représentation est également l'occasion d'une réflexion sur la qualité des graphiques résultants. La conclusion dresse le

Visualisation par l'exemple des dépendances dans les bases de données relationnelles

Fabien De Marchi*, Jean-Marc Petit**

*Laboratoire LIRIS, UMR CNRS 5205
Université Claude Bernard - Lyon 1
8, boulevard Niels Bohr, 69 622 Villeurbanne cedex France
fabien.demarchi@liris.cnrs.fr
**Laboratoire LIMOS, UMR CNRS 6158
Université Blaise Pascal - Clermont-Ferrand II
24 avenue des Landais, 63 177 Aubière cedex, France
jmpetit@math.univ-bpclermont.fr

Résumé. Comprendre la sémantique des bases de données relationnelles existantes est important pour de nombreuses applications. Cette sémantique est principalement véhiculée par les dépendances fonctionnelles (DF) et les dépendances d'inclusion (DI) ; elles généralisent respectivement les notions de clé et de clé étrangère. Toutefois, il est fréquent que les bases de données opérationnelles deviennent désordonnées dans le temps ; dans ce cas, les contraintes d'intégrité doivent être retrouvées à partir des données. Plusieurs méthodes ont été proposées pour la découverte des DF ou des DI. Ces algorithmes fournissent à l'administrateur un ensemble de dépendances satisfaites dans les données.

Se pose alors le problème de la compréhension des dépendances extraites, incluant des aspects liés à la visualisation des connaissances. Cette étape doit permettre, par exemple, d'assister l'utilisateur final à sélectionner les règles intéressantes, ou à comprendre pourquoi une dépendance attendue n'est pas satisfaite dans les données. Nous proposons de fournir à l'administrateur ou l'analyste, en complément de la liste des règles, un échantillon de la base de données, vérifiant exactement les même DF et DI, appelé *base de données d'Armstrong informative* (BDAI). Ces exemples nous semblent particulièrement adaptés pour faciliter les échanges entre l'administrateur et les experts du domaine. Nous donnons certaines propriétés sur l'existence et la taille des BDAI, ainsi que des algorithmes pour les construire. Des expérimentations sur une base réelle issue du web montrent l'intérêt pratique de cette proposition.

1 Introduction

Comprendre la sémantique des bases de données relationnelles existantes est important pour de nombreuses applications. Parmi elles, citons des travaux de rétro-conceptions [Casanova et de Sa, 1983, Markowitz et Makowsky, 1990, Petit *et al.*, 1996, Comyn-Wattiau et Akoka, 1999], dont le but est de retrouver le schéma conceptuel des données à partir de leur forme relationnelle, des travaux sur l'intégration de données

Découverte interactive de règles d'association via une interface visuelle

Pascale Kuntz, Rémi Lehn
Fabrice Guillet, Bruno Pinaud

Laboratoire d'informatique de Nantes Atlantique (LINA)
Site Ecole Polytechnique
La Chantrerie - rue Christian Pauc
BP 50609, 44306 Nantes Cedex 3
{pascale.kuntz, remi.lehn, fabrice.guillet, bruno.pinaud}@univ-nantes.fr,

Résumé. En nous appuyant sur des hypothèses majoritairement empruntées à des travaux sur les systèmes anthropocentrés d'aide à la décision, nous décrivons dans cet article un environnement interactif de fouille de règles d'association dans lequel l'utilisateur pilote le processus, en jouant le rôle d'une heuristique dans un environnement de recherche complexe. Afin de permettre à la fois une représentation visuelle accessible et une instanciation aisée des outils d'interactivité le modèle choisi est ici un graphe en niveaux - les niveaux étant associés aux cardinaux des sous-ensembles d'attributs des prémisses. Le processus a été déployé dans un logiciel prototype dont l'analyse des résultats ouvre de nouvelles perspectives sur l'analyse comportementale d'un utilisateur en situation de fouille.

1 Introduction

Si les efforts les plus remarqués de la première décennie de l'Extraction de Connaissances dans les Données (ECD) ont principalement porté sur le développement d'algorithmes automatiques performants, le rôle de l'utilisateur est peu à peu devenu un sujet de préoccupation majeur. Ce besoin d'intégration a conduit à l'émergence de nombreux outils de visualisation et d'interaction [Fayyad *et al.*, 2002], [Grinstein, 1996]. Il s'agit d'apporter à l'utilisateur un substrat artificiel qui transcrive un grand nombre d'informations et qui soit un support à ses connaissances et à son intuition pour lui permettre de découvrir de nouvelles relations et d'imaginer de nouvelles questions.

1.1 Visualisation et interaction en ECD

Traditionnellement, le recours à des techniques de visualisation précède et suit les étapes de traitement automatiques. En amont du processus d'ECD, elles assistent les tâches de sélection et de pré-traitement des données en renforçant la convivialité des interfaces. En aval, à l'issue de l'application des algorithmes de découverte de structures, elles visent à présenter des résultats sous des formes intelligibles facilitant leur interprétation. Dans ce contexte, la plupart des outils ont été définis pour des objectifs spécifiques (recherche de règles d'association, classification de données spatiales, etc.).

Représentation et comparaison de séquences par visualisation

Christine Largeron (*), Cedric Dreissia (**)

Université Jean Monnet de Saint-Etienne
(*) EURISE
23, rue du docteur Paul Michelon
(**) CREUSET
6, rue Basse des Rives
42023 Saint-Etienne Cedex 2
Christine.Largeron@univ-st-etienne.fr

Résumé. Dans cet article, nous présentons un outil de visualisation de séquences modélisées par des arbres de suffixes probabilistes (Prediction suffix trees - PST). Ce type d'arbre permet de représenter une chaîne de Markov d'ordre variable. Dans différentes applications, il s'est avéré plus efficace qu'une chaîne de Markov d'ordre fixe, avec un coût calculatoire moindre. Pour ces raisons, il nous a paru intéressant d'exploiter le caractère arborescent de ce mode de représentation des séquences, non seulement d'un point de vue algorithmique, mais aussi d'un point de vue visuel. Le logiciel que nous avons développé dans ce but fournit une représentation graphique d'un PST appris à partir de séquences et, il permet de le comparer à un autre. Dans un contexte de classement supervisé d'une nouvelle séquence, il apporte une information complémentaire par rapport au PST en mettant en évidence les sous-séquences qui n'ont pas été observées dans la nouvelle séquence bien qu'elles soient caractéristiques du modèle sous-jacent à sa classe d'affectation. Ainsi, il permet de mieux appréhender la structure des séquences et d'améliorer le processus de fouille de données par leur visualisation.

1 Introduction

Les travaux précurseurs en fouille visuelle de données (Visual Data Mining) remontent à Bertin ou encore à Tufte [Bertin, 1977, Tufte, 1983]. Ils portaient sur la représentation graphique de données. Jusqu'à un passé proche, les techniques de visualisation étaient principalement employées dans deux étapes lors du processus de traitement de données :

- au début de la chaîne du traitement, dans une phase exploratoire des données brutes,
- à la fin du traitement, dans une phase de présentation des résultats sous une forme souvent plus synthétique.

Avec l'émergence de la fouille visuelle de données [Card *et al.*, 1999, Spence, 2001, Keim, 2002, Davidson et Soukup, 2002, Poulet, 2004], elles interviennent dans la phase principale du processus de fouille, afin d'impliquer plus directement l'utilisateur

Détection et interprétation visuelle d'outliers dans les grands ensembles de données

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 hybride de détection d'outliers (individus atypiques) dans de grands ensembles de données, utilisant un algorithme génétique pour la sélection des attributs et une approche basée sur la distance pour la détection de l'élément outlier (atypique) suivant ce sous-ensemble d'attributs. Une fois l'outlier trouvé, nous essayons de l'expliquer : est-ce une erreur, un bruit ou une valeur significativement différente des autres ? Pour ce faire, on utilise des méthodes visuelles telles que les coordonnées parallèles. Nous évaluons les performances de notre méthode sur différents ensembles de données de grandes dimensions et le comparons avec les algorithmes existants.
Mots-clefs. fouille de données, détection d'outliers, visualisation, algorithmes génétiques, coordonnées parallèles, grands ensembles de données.

1. Introduction

Le développement du réseau internet et la baisse des coûts du matériel informatique ont permis à de nombreux organismes de constituer de grandes masses de données trop volumineuses et complexes pour pouvoir être appréhendées par un utilisateur. L'Extraction de Connaissances à partir de Données (ECD) est née de ce besoin, on la définit comme étant l'extraction de nouvelles connaissances potentiellement utiles à partir de grandes quantités de données [Fayyad et al. 1996], le cœur du processus d'ECD est la fouille de données (Data Mining). Dans ce cadre précis (de fouille de données), nous nous intéressons à la recherche d'outliers (individus atypiques). La recherche d'outliers a de nombreuses applications telles que la détection de fraudes, la recherche pharmaceutique, les applications financières, le marketing, etc. Un outlier (individu atypique) est un petit ensemble de données, un point ou une observation qui est considérablement différent, divergent, dissemblable ou distinct du reste des données. Le problème est alors de définir cette dissimilitude entre objets, ce qui caractérise un outlier. Typiquement, celle-ci est estimée par une fonction calculant la distance entre objets, la tâche suivante consiste à déterminer les objets les plus éloignés de la masse. Certaines difficultés apparaissent lorsque l'on est face à des ensembles de données ayant un grand nombre de dimensions en terme d'attributs. En effet, dans les ensembles à grandes dimensions, les données sont rares et la notion de voisinage perd de son sens. La rareté dans les espaces de grandes dimensions implique que tout point est candidat pour être un bon outlier et donc la recherche d'outliers devient complexe et coûteuse en temps de calcul.

La plate-forme DynaSpat : Les Dynamiques Spatiales

Sandrine Coelho*, Christine Thomas-Agnan**,
Nicolas Lassabe***, Yves Duthen****

Université des Sciences Sociales (UT1)-Manufacture des Tabacs
21, Allée de Brienne -31000 Toulouse
*sandrine.coelho@univ-tlse1.fr
Entreprise GéoSignal
**cthomas@cict.fr
Laboratoire GREMAQ
***lassabe@irit.fr
Laboratoire IRIT-UT1, entreprise GéoSignal
****duthen@irit.fr
Laboratoire IRIT-UT1

Résumé. La plate-forme DynaSpat[1] est une plate-forme de simulation et de visualisation pour la description, l'analyse et la modélisation de comportements d'acteurs économiques. Elle intègre un ensemble de logiciels et de bibliothèques permettant d'exécuter des simulations basées sur les algorithmes génétiques et de résumer une information complexe, issue d'un jeu de données, par des variables pertinentes. Elle permet ainsi de comprendre l'influence réciproque entre la structuration d'un territoire et des comportements d'acteurs économiques ou sociaux.

1. Introduction

L'étude des dynamiques spatiales s'applique à des données complexes et hétérogènes provenant de sources multiples, il est donc indispensable de disposer d'outils permettant de les interpréter et d'en retirer des connaissances. Pour cela, il faut maîtriser plusieurs disciplines telles que les sciences et technologies de l'information, l'économie, la sociologie, la statistique et la géographie.

Nous pouvons constater que bien des outils sont développés en ce qui concerne le stockage, l'extraction et la visualisation de données. Ils sont optimisés pour leur domaine d'application mais, de ce fait, leur utilisation est parfois trop spécifique. De plus, on peut remarquer que bien souvent le couplage de ces modules peut s'avérer difficile et que les outils d'analyse disponibles s'avèrent insuffisants. Notre approche consiste à réunir et compléter ces techniques dans un environnement unique.

L'objectif du projet DynaSpat est d'expliquer l'émergence, dans une zone géographique donnée, de comportements spécifiques d'acteurs économiques, tout en permettant la convergence de plusieurs champs disciplinaires. Nous avons créé un outil global permettant d'utiliser un outil d'optimisation (AGMC) et de coupler une carte avec un ensemble de techniques statistiques spatiales (GéoXP) dialoguant de façon interactive. Le principal apport

[1] Remerciements : ce projet de recherche est supporté par la région Midi-Pyrénées et regroupe trois laboratoires : IRIT, GREMAQ, LEREPS, et l'entreprise GéoSignal.