

Une approche MDS hybride pour l'exploration visuelle interactive

Fabien Jourdan, Guy Melançon

LIRMM
161, rue Ada
34392, Montpellier, France
{fjourdan, melancon}@lirmm.fr

Christophe Douy, Alexandre Gasne

Pikko Software
200, impasse des Pâquerettes
34170, Castelnau-le-Lez, France
{cdouy, alex}@pikko-software.com

RESUME

L'article décrit une méthode permettant d'explorer un espace d'information dont les éléments sont associés à une taxonomie (une classification d'un domaine d'activité). On dispose aussi d'un vecteur d'attributs numériques pour chaque élément. Le cœur de la méthode repose sur l'utilisation d'un algorithme de plongement cherchant à placer à proximité les éléments « similaires ». Le caractère dynamique de l'algorithme en fait un outil adapté à l'exploration interactive. Nos premiers résultats présentés ici, et appliqués à un ensemble d'entreprises régionales, confirme notre méthode comme un outil pertinent pour la recherche d'opportunité de fusion d'entreprises selon diverses caractéristiques économiques.

MOTS CLES : Visualisation, exploration interactive, MDS, taxonomie, classification.

ABSTRACT

The paper presents a novel technique for the exploration of an information space where elements are associated with a taxonomy and where each element has an associated attribute vector. A hybrid MDS method is at the core of the method. The underlying algorithm embeds elements in the 2D plane, trying to place "similar" elements close to one another. The dynamic character of the algorithm makes it a well suited tool for interactive exploration. Our first results presented here confirm our approach as a tool relevant for finding merging opportunities between companies based on various economic characteristics.

CATEGORIES AND SUBJECT DESCRIPTORS: H.4 [Information Systems Applications]: I.3.6 [Computer Graphics]: Methodology and Techniques – Interaction Techniques.

GENERAL TERMS: Algorithms, Management

KEYWORDS: Visualization, interactive exploration, MDS, taxonomy, classification.

INTRODUCTION

La méthode présentée ici, et correspondant aux premiers résultats d'un travail en cours, apporte une aide à la recherche d'opportunités de fusion entre petites et moyennes entreprises régionales. Par examen des profils des entreprises (domaine d'activité, taux d'activité, chiffre d'affaires, etc.), l'analyste doit pouvoir proposer des scénarios de fusion d'entreprises qu'il doit pouvoir justifier auprès des acteurs impliqués et des tutelles.

Nous avons cherché à offrir un outil où l'ensemble des entreprises est visualisé, et où les entreprises de profils « similaires » sont placées à proximité à l'écran. Nous proposons une méthode qui permet à l'utilisateur de librement déplacer les icônes représentant les entreprises et d'influer sur l'animation de la cartographie des entreprises, dans le but de l'aider dans sa réflexion.

L'idée centrale réside dans l'utilisation d'une méthode de plongement capable de calculer un bon placement des (icônes représentant les) entreprises sur la base de leurs *dissimilarités* : une valeur numérique mesurant à quel point les profils de deux entreprises sont lointains. Nous avons adapté la méthode afin qu'elle prennent aussi en compte une information additionnelle. Outre les divers paramètres économiques (des attributs numériques), nous disposons d'une *taxonomie* décrivant le ou les domaines d'activités d'une entreprise sélectionnés parmi une classification de ces domaines et sous-domaines.

Nous allons d'abord présenter la méthode de plongement MDS pour ensuite décrire l'adaptation que nous en proposons afin de prendre en compte la classification des domaines d'activités.

MDS (MULTI-DIMENSIONAL SCALING)

Le problème de départ de la théorie du « multi-dimensional scaling » consiste à plonger un ensemble d'éléments $X = \{x_1, \dots, x_N\}$ dans un espace euclidien de manière à ce que la distance euclidienne entre ces éléments se rapproche au mieux de *dissimilarités* entre les éléments deux à deux (on notera par δ_{ij} la dissimilarités entre les éléments i et j). La possibilité de calculer un plongement qui soit satisfaisant dépend grandement des

propriétés de la matrice de dissimilarités elle-même. Outre le problème qui consiste à calculer un plongement, la littérature se penche sur les questions concernant la matrice de dissimilarités, et aussi sur l'estimation de la qualité des plongements [10], [6], [1].

Les approches spectrales, parmi les plus anciennes permettant de calculer un plongement reposent sur l'algèbre linéaire, exploitent les propriétés de la matrice de dissimilarités. Des approches plus récentes misent sur un calcul par approximation basé sur des analogies physiques. Les éléments de données à plonger dans l'espace euclidien sont assimilés des corps physiques, et les forces agissant sur eux sont fonction des dissimilarités. En quelques mots, on peut calculer pour chaque paire d'éléments une force qui permet de déplacer les sommets, les repoussant l'un l'autre lorsqu'ils sont à distance moindre que la dissimilarité cible, les rapprochant en cas contraire. Le plongement lui-même correspond donc à *simuler* le système physique, en misant sur sa convergence vers un état stable. Des travaux récents ont apportés des améliorations permettant d'utiliser ces algorithmes en situation d'interaction, alors que l'approche naïve simulant le système physique est en complexité $O(N^3)$ (où N est le nombre d'éléments à plonger dans l'espace) [2], [7], [8], [5]. Munzner a développé une approche originale permettant à l'utilisateur d'intervenir et de guider interactivement le processus de plongement [11]. L'algorithme de plongement que nous avons utilisé utilise les versions plus récentes calculant la simulation en temps $O(N \log N)$ [5].

L'approche MDS peut typiquement être utilisée lorsque les éléments de données sont associés à des vecteurs d'attributs numériques $v_i = (a^1_i, \dots, a^p_i)$, comme dans notre cas. En effet, on peut alors poser $\delta_{ij} = \|v_i - v_j\|$ (distance euclidienne entre les vecteurs d'attributs). C'est le cas que nous décrivons un peu plus loin : les entreprises pourront être placées les unes par rapport aux autres en prenant comme dissimilarités la distance qui sépare les vecteurs d'attributs qui leurs sont associés. On peut alors s'attendre à ce que deux entreprises placées à proximité par l'algorithme aient des « profils » similaires ».

UNE APPROCHE MDS HYBRIDE

Le cas qui nous intéresse est celui où les éléments de données sont de plus associés à certains concepts d'une taxonomie. Il faut penser à la taxonomie comme à une description des catégories d'un domaine d'activité (ou domaine métiers) implicite au contexte des données.

Cette taxonomie est habituellement décrite par une structure arborescente où le schéma arborescent correspond à un affinage des catégories. Nous allons d'abord considé-

rer le cas très simple d'une taxonomie consistant en une simple description de catégories sans sous-catégories. Dans ce cas particulier, l'arbre est réduit à une racine possédant k fils. On plonge cette taxonomie dans le plan en associant à chacun des fils un arc de cercle qui s'étend sur un angle de $2\pi/k$, résultant en un motif circulaire. En réalité, les fils de la racine ne sont soumis à aucun ordre, alors que ce plongement introduit de fait un ordonnancement circulaire. La figure 1 illustre ce procédé.

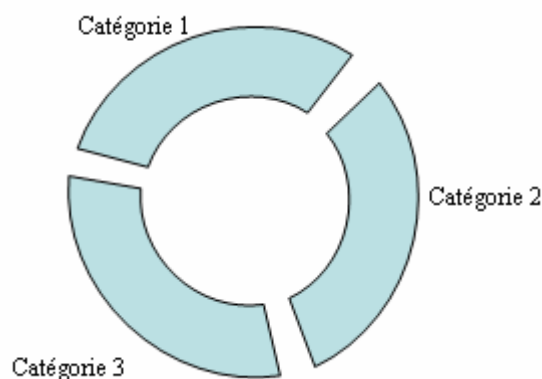


Figure 1. Disposition circulaire des catégories d'une taxonomie.

Afin de s'assurer que les éléments de données sont placés à proximité des catégories auxquelles ils sont associés, on introduit dans la simulation des forces d'attraction agissant sur eux. Les catégories sont quand à elle fixées une fois pour toute dans la simulation. Bien entendu, les bandes circulaires ne sont pas affichées mais sont seulement présentes au niveau du modèle interne.

La figure 2 montre un instantané de l'algorithme sur un petit ensemble d'éléments de données. Notez qu'il s'agit bien d'un instantané puisque les bénéfices de l'algorithme résident en sa capacité à réagir aux déplacements des éléments effectués par l'utilisateur, provoquant une perturbation momentanée. La simulation, animée, ramène ensuite le système à un état stable. Les mouvements et déplacements mutuels des éléments facilitent la lecture de la carte et l'identification de motifs structuraux. En effet, ces mouvements introduits par l'utilisateur sont parfois aptes à « débloquer » des situations permettant à un élément particulier de rejoindre un endroit qui lui était jusqu'alors inaccessible. C'est sur ces constats que nous misons pour aider l'utilisateur dans sa tâche.

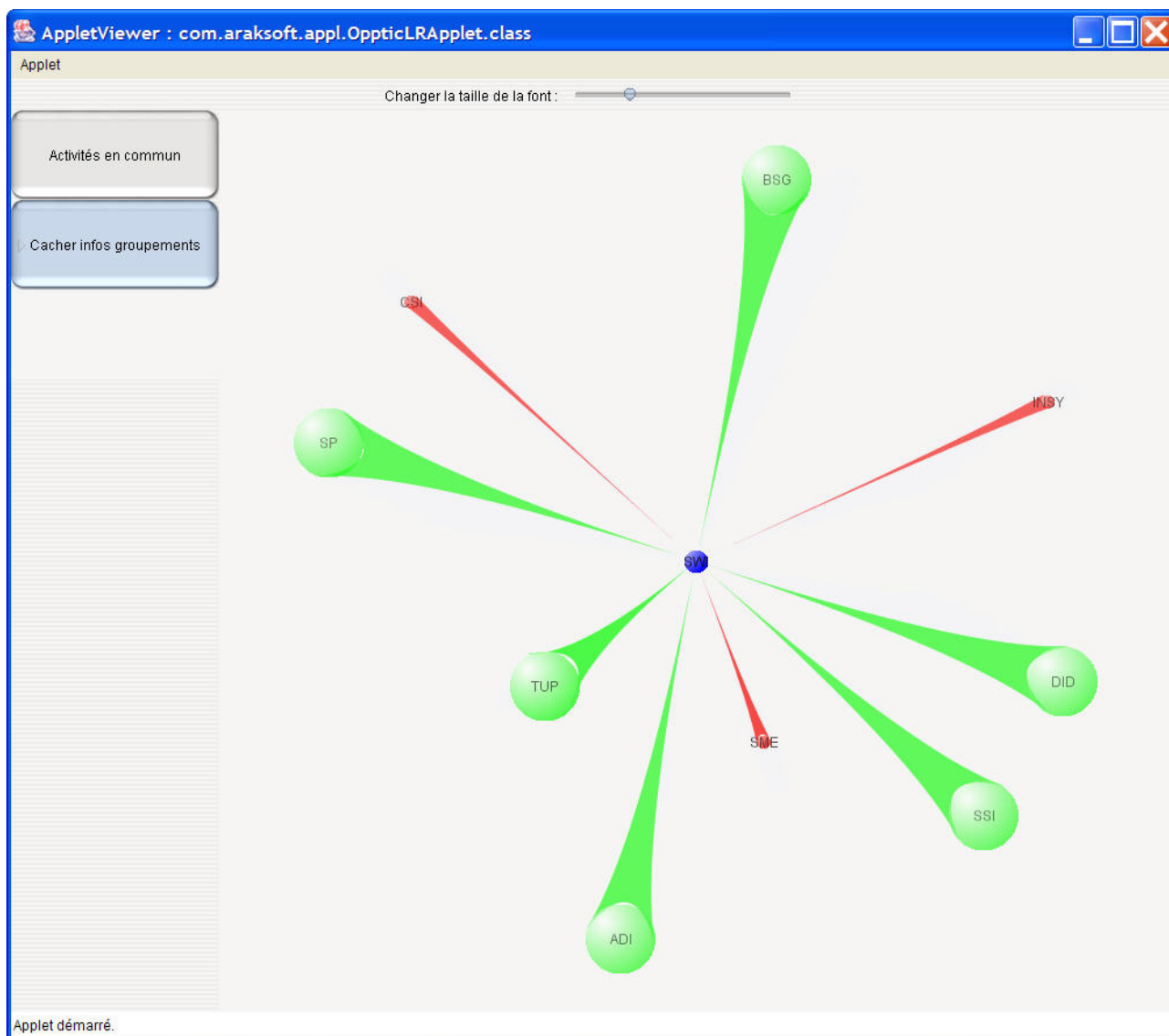


Figure 2. L'un des trois modes de navigation de l'interface permet de saisir en un coup d'oeil la position relative d'une entreprise dans le réseau régional.

Dans l'exemple illustré par la figure 2, l'utilisateur est à même de comparer la place d'une entreprise (en bleu, au centre de la fenêtre) dans le réseau régional. La distance des entreprises voisines traduit leurs affinités avec l'entreprise étudiée et se calcule en fonction d'une série de mesures reflétant la santé des entreprises, leur niveau d'activité ou encore leur taille; la position relative des entreprises en périphérie indique aussi des rapprochements possibles entre elles. La coloration est induite des attributs et confirme ou informe les opportunités de fusion.

LE CAS D'UNE TAXONOMIE PLUS RICHE

Le cas d'une taxonomie qui ne se réduit pas à un ensemble de catégories peut être pris en compte. Observez d'abord que si un élément est associé à une catégo-

rie 1.1, sous-catégorie de la catégorie 1, alors il est implicitement aussi associé à la catégorie 1. Cette idée justifie que les sous-catégories induisent un découpage de l'arc de cercle associé à la catégorie mère. La figure 3 illustre cette construction. Le procédé peut être répété en principe un nombre arbitraire de fois, mais il faut toutefois avouer que la lisibilité s'en trouve rapidement pénalisée. La simulation doit elle aussi être modifiée de manière à pouvoir induire une séparation entre les éléments de données qui sont soit associés à des sous-catégories différentes.

Cette idée apporte une aide réelle à l'utilisateur qui peut interactivement « ouvrir » ou « fermer » les catégories et sous-catégories de la taxonomie au gré de la navigation.

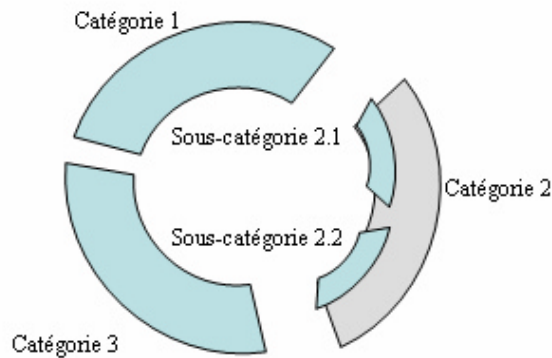


Figure 3. Les sous-catégories découpent la bande correspondant à la catégorie mère.

PERSPECTIVES

La méthode présentée ici a été utilisée avec satisfaction par nos premiers utilisateurs. Il faut souligner que l'algorithme de plongement n'apporte pas à lui seul une visualisation répondant aux besoins des analystes, et que c'est son caractère dynamique qui est son atout. Cette dynamique tient à la fois à la possibilité d'animer les mouvements des éléments de données au cours de la simulation, mais aussi à la possibilité de raffiner la taxonomie au gré de l'exploration.

Distorsion appliqué à la taxonomie

Cet affinage de la taxonomie pourrait être jumelé à un effet fish-eye que nous projetons de développer. En effet, dans la suite des idées introduites par Furnas [4] et poursuivies par exemple par Schaffer [9], l'intérêt de l'utilisateur pour une catégorie particulière justifie que l'on consacre un espace plus conséquent aux éléments relevant de cette catégorie. Dans notre cas, cette transformation pourra être facilement implémentée en induisant un étirement de la bande correspondant à la catégorie sélectionnée (accompagnée d'une compression égale sur les autres catégories).

Extension aux ontologies

Parmi les premières réactions à l'outil que nous avons développé apparaît clairement la demande de pouvoir travailler avec une description des entreprises à l'aide d'un réseau de concepts (une ontologie). Plusieurs avenues sont à explorer, notamment le travail de Fluit *et al.* [3].

Tests utilisateurs

Nos contacts avec les utilisateurs finaux sont pour l'instant restés informels. Une étude plus attentive et plus structurée devrait nous permettre de gagner du recul par rapport à ce premier outil, mais aussi de confirmer la possibilité d'aller au-delà de simples taxonomies. En effet, sans préjuger de la valeur des ontolo-

gies, il se pourrait que leur complexité pose certaines difficultés au niveau de la représentation et la rende moins lisible. Cette question est évidemment à étudier de près.

BIBLIOGRAPHIE

1. Borg, I. and P. Groenen (1997). *Modern Multidimensional Scaling: Theory and Applications*, Springer Verlag.
2. Chalmers, M. (1996). A Linear Iteration Time Layout Algorithm for Visualizing High-Dimensional Data. IEEE Symposium on Information Visualization, San Francisco, USA, IEEE Computer Society, pp. 127-132.
3. Fluit, C., M. Sabou, and F. v. Harmelen, 2002: *Ontology-based Information Visualisation*. Visualising the Semantic Web, V. Geroimenko, Ed., Springer Verlag, 2002.
4. Furnas, G. W., 1986: Generalized Fisheye Views. *Human Factors in Computing Systems CHI '86*, ACM Press, 16-23.
5. Jourdan, F. and G. Melançon (2004). Multiscale Hybrid MDS. IV 2004, 8th International Conference on Information Visualization, London, UK, IEEE Computer Society, pp. 388-393.
6. Kruskal, J. B. and M. Wish (1978). Multidimensional Scaling, Sage Publications.
7. Morrison, A., G. Ross, et al. (2002). A Hybrid Layout Algorithm for Sub-Quadratic Multidimensional Scaling. IEEE Symposium on Information Visualization, Boston, USA, IEEE Computer Society, pp. 152-158.
8. Morrison, A. and M. Chalmers (2003). Improving Hybrid MDS with Pivot-Based Searching. IEEE Symposium on Information Visualization, Seattle, USA, IEEE Computer Society, pp. 85-90.
9. Schaffer, D., Z. Zuo, S. Greenberg, L. Bartram, J. Dill, S. Dubs, and M. Roseman, 1996: Navigating Hierarchically Clustered Networks through Fisheye and Full-zoom Methods. *ACM Transactions on Computer-Human Interaction*, 3, 162-188.
10. Torgerson, W. S. (1952). "Multidimensional Scaling: Theory and Method." *Psychometrika* 17: 401-419.
11. Williams, M. and T. Munzner, 2004: Steerable, Progressive Multidimensional Scaling. IEEE Symposium on Information Visualization, Austin, Texas, IEEE Computer Society, 57-64.