La prise en compte de la dimension temporelle dans la classification de données

Eloïse Loubier *, Bernard Dousset *

* I.R.I.T. (Institut de Recherche en Informatique de Toulouse),
118 route de Narbonne, 31062 TOULOUSE Cedex 9
{loubier, dousset}@irit.fr;

Résumé. Dans un contexte d'ingénierie de la connaissance, l'analyse des données relationnelles évolutives est une question centrale. La représentation de ce type de données sous forme de graphe optimisé en facilite l'analyse et l'interprétation par l'utilisateur non expert. Cependant, ces graphes peuvent rapidement devenir trop complexes pour être étudiés dans leur globalité, il faut alors les décomposer de manière à en faciliter la lecture et l'analyse. Pour cela, une solution est de les simplifier, dans un premier temps, en un graphe réduit dont les sommets représentent chacun un groupe distinct de sommets : acteurs ou termes du domaine étudié. Dans un second temps, il faut les décomposer en instances (un graphe par période) afin de prendre en compte la dimension temporelle.
La plateforme de veille stratégique Tétralogie, développée dans notre laboratoire, permet de synthétiser les données relationnelles évolutives sous forme de matrices de cooccurrence 3D et VisuGraph, son module de visualisation, permet de les représenter sous forme de graphes évolutifs.
VisuGraph assimile les différentes périodes à des repères temporels et chaque sommet est placé en fonction de son degré d'appartenance aux différentes périodes. Ce prototype est aussi doté d'un module de la classification interactive de données relationnelles basé sur une technique de Markov Clustering, qui conduit à une visualisation sous forme de graphe réduit. Nous proposons ici de prendre en compte la dimension temporelle dans notre processus de classification des données. Ainsi, par la visualisation successive des différentes instances, il devient plus facile d'analyser l'évolution des classes au niveau intra mais aussi au niveau inter classes.

1 Introduction

L'étude de la migration des termes, en particulier de l'évolution des données relationnelles issues de la synthèse de grands corpus d'information est un aspect majeur dans l'ingénierie de la connaissance et en particulier dans le cadre de la veille. Dans ce contexte, le recours à la visualisation de données par des graphes apporte un réel confort aux utilisateurs, qui, de façon intuitive, peuvent s'approprier une forme de connaissance difficile à décrire autrement. Bien souvent, ces graphes sont trop complexes pour être étudiés dans leur globalité, il faut alors les décomposer de manière à faciliter la lecture et l'analyse des données. Une première simplification du graphe est réalisé par le biais de la classification en un graphe réduit dont les sommets représentent chacun un groupe distinct d'acteurs ou de

termes du domaine. D'autre part, la décomposition en graphes de périodes simplifie la structure de la représentation, en prenant en compte la dimension temporelle.

Dans ce contexte, le prototype VisuGraph, module de la plate-forme de veille stratégique Tétralogie, permet déjà la visualisation de données évolutives. Basée sur une visualisation globale, toutes périodes confondues, puis individuelle, les données sont représentées sous forme de sommet dont les coordonnées traduisent les caractéristiques temporelles.

Ce prototype est aussi doté de la classification interactive de données relationnelles, basée sur une technique de Markov Clustering, qui permet à la fois d'obtenir des classes homogènes et le graphe réduit dont les sommets sont les classes obtenues. Nous l'avons aménagée pour pouvoir intervenir sur le nombre de classes (augmentation, diminution), mais celui-ci reste assez aléatoire. Nous proposons alors de prendre en compte la dimension temporelle afin d'analyser l'évolution des différentes classes de données obtenues par l'algorithme MCL au cours du temps. Cette technique prend en compte la topologie du graphe (diamètre, centralité, adjacence, flux, …) dans un contexte évolutif, tout en s'appliquant à une métrique algébrique classique : la distance.

Dans la première partie de cet article, la plate forme de veille Tétralogie est présentée, en mettant l'accent sur l'extraction et le traitement des données. Puis, le module de représentation graphique VisuGraph est proposé. Dans la section 3, nous détaillons notre proposition, en développant l'analyse évolutive des classes de données, que nous expérimentons. Enfin, dans la dernière section, nous concluons sur nos travaux et présentons nos perspectives.

2 Présentation de Tétralogie

Tétralogie (Dousset et al., 1988) est un logiciel de macro-analyse de données textuelles semi-structurées intégrant la dimension temporelle. Les données analysées par la plateforme Tétralogie sont issues de bases de données, de revues, de journaux, de périodiques, de revues de veille technologique, des thèses et de brevets ou encore de CDROMS. Les informations extraites de ces sources sont synthétisées sous forme de matrices de cooccurrence, exploitables dans les différents modules proposés par Tétralogie.

Les unités de base de toute analyse sont le terme, le champ (auteur, mots-clefs, adresse, date, ...) et le document. Un champ est une balise prédéfinie de la base de donnée semi-structurée, par exemple auteur, date, adresse, organisme. Un champ, peut être mono-valué (journal) ou multi-valué (auteur, mot-clef,…). Un terme est une unité textuelle correspondant au contenu d'un champ mono-valué ou une partie d'un champ multi-valué délimité par des séparateurs. Les données analysées peuvent être croisées entre deux champs, sous champs ou groupes de champs afin d'obtenir des matrices de fréquence, de présence/absence ou encore de co-occurrence (une des variables peut contenir plusieurs champs : auteurs, mots, clés…) sur lesquelles porteront ensuite les analyses. Dans le cas de données évolutives, un champ relatif au temps est pris en compte. Dans un contexte temporel, on aura autant de matrices de croisement que de périodes (appelées aussi « instances ») étudiées. Chaque croisement au sein d'une matrice est appelé « valeur de métrique » d'un ou plusieurs champs. Certaines informations sont sémantiquement équivalentes ou hiérarchisées, aussi est-il très utile de disposer de la fonctionnalité de Tétralogie permettant de radicaliser les données en les transformant en données simples, de les nettoyer mais aussi de normaliser ou d'homogénéiser les termes (adresse, organisme, date), tout en contrôlant qu'il n'y ait pas d'ajout ou encore d'oubli de caractère au sein de chaque terme (Loubier et al., 2007).

3 La visualisation évolutive de données relationnelles

3.1 La prise en compte de la dimension temporelle

Le prototype VisuGraph est un module de visualisation, offrant à la plateforme Tétralogie la possibilité de représenter les données matricielles sous forme de graphe. Dans cet article, nous ne traitons que des graphes non orientés. Considérons le graphe non orienté G = (S,A) où S est l'ensemble fini des éléments appelés sommets ou encore nœuds. A est l'ensemble fini des liens appelés arêtes, liant les sommets. $\mathbf{A \subset S \times S = \{(s,t) \mid s,t \in S\}}$
Chaque sommet est assimilé à la valeur de la métrique d'un seul champ alors que chaque lien correspond à la valeur du croisement de deux champs. Afin d'obtenir un graphe planaire, sans qu'aucune arête n'en croise une autre, nous avons recours à l'algorithme « force-directed placement » (FDP) (Eades, 1984), assimilant les sommets à des masses et chaque arête d'un graphe à un ressort reliant les sommets. Un tel système engendre des forces entre les sommets, ce qui entraine des déplacements respectifs. Après une phase de transition le système se stabilise. La condition d'arrêt est un nombre maximum d'itérations.

Dans ce cas d'analyse temporelle, un graphe global représente toutes les données, toutes périodes comprises, puis chaque graphe de période est visualisé individuellement, réalisant ainsi une animation. Dans ce contexte évolutif, nous attribuons un sommet virtuel (non visible dans le dessin mais dont la présence est prise en compte dans le graphe) qui servira de repère pour chaque période considérée. Ces sommets virtuels sont fixés dans un ordre chronologique et de façon équidistante sur le contour de la fenêtre de visualisation (comme les heures sur un cadran) (Loubier2, 2007). Le dessin de graphe est influencé par l'attribution de nouveaux arcs reliant chacun des sommets aux repères temporels, qui le concernent, en leur attribuant un poids plus important que la valeur de la métrique d'arête la plus grande. Ceci engendre un déplacement, vers certains repères, en fonction de la plus ou moins forte présence d'un sommet dans chaque période.

3.2 Simplification sous forme de graphe réduit

Afin de faciliter l'analyse, les données les plus fortement liées doivent être regroupées en classes homogènes. Parmi les travaux effectués sur le partitionnement de graphe, (Alpert et al. 1995, Kuntz et al. 2000, Jouve et al. 2001) se basent sur des approches spectrales alors que les algorithmes de la famille METIS (Karypis et al. 1998) se basent sur le partitionnement multi niveaux. La méthode de partitionnement utilisée dans VisuGraph est inspirée du Markov Clustering (Van Dongen 2000) que nous avons aménagée pour pouvoir influencer le nombre de classes proposées (Karouach, 2003). Cette approche calcule des probabilités de transition entre tous les sommets du graphe en partant de la matrice de transition des marches aléatoires. Deux simples opérations matricielles sont successivement itérées. La première calcule les probabilités de transition par des marches aléatoires de longueur fixée r et correspond à une élévation de la matrice à la puissance r. La seconde consiste à amplifier les différences en augmentant les transitions les plus probables et en diminuant les transitions les moins probables. Les transitions entre sommets d'une même communauté sont alors favorisées et les itérations successives des deux opérations conduisent à une situation limite dans laquelle seules les transitions entre sommets d'une même communauté sont possibles.

Soient un graphe G = (V,w) où w : V x V → R^+ et M_G la matrice associée à G. T_G est la matrice normalisée (somme des poids des arcs sortants = 1). Soit T_{pq} la probabilité d'effectuer la transition p → q(T = T_G). L'expansion s'effectue par multiplication des matrices, visant à élargir la capacité de l'arc entre deux nœuds. L'inflation d'une colonne vise la promotion des voisins favoris au détriment de ceux moins favoris.

Pour r>0, $(\Gamma,T)_{pq} = \frac{(T_{pq})^r}{\sum_{i=1}^{k} (T_{iq})^r}$

```
Algorithme de MCL (G, [e_i]_{i∈N}, [r_i]_{i∈N})
  T1 ← T_G;
  k ← 0;
Tant que T 2k+1 n'est pas (approx.) idempotente
  k← k+1 ;
  T_2k ← Exp_ek(T_2k-1) ;
  T_2k+1 ← Γr_k (T_2k);
```

L'évaluation de la méthode MCL a montré la rapidité et la qualité de ses résultats (Enright et al. 2002). Le graphe obtenu est alors un graphe de classe, pour lequel chaque sommet représente une classe. Les liens entre les sommets sont assimilés aux liaisons interclasses. Dans un second temps, l'attribution d'une couleur spécifique à chaque classe permet de visualiser le graphe complet, en figeant un représentant par classe et en distribuant les autres sommets sur une couronne centrée sur ce dernier, permettant ainsi une vue intra classe.

4 Expérimentation

Le graphe global (au centre de la FIG 1) sert de base lors de l'analyse. C'est sur ce dernier que s'effectue le partitionnement des données, sans prendre en compte les arcs liants les sommets aux repères temporels, afin de conserver les propriétés classiques de classification. Une fois le graphe réduit obtenu, les liens entre sommets et repères temporels sont instaurés et l'application de l'algorithme FDP permet de placer chaque représentant de classe selon ses caractéristiques temporelles. La visualisation successive de chaque graphe de période permet de les comparer, au niveau de leur structure inter classes mais aussi au niveau des sommets composant les différentes classes (intra classe). L'expérimentation effectuée se base sur un corpus portant sur les auteurs ayant publié un article lors des quatre dernières sessions du colloque VSST. Il est important de noter que toutes les classes sont reliées pour l'ensemble des quatre périodes visualisées simultanément, laissant penser que le graphe est connexe. La représentation par période des données montre que les données ne sont pas forcément toutes reliées entre elles pour chaque tranche de temps, révélant ainsi la non connexité du graphe initial. Par masquage des données du graphe global n'appartenant pas à la première tranche de temps, le graphe de la première période est obtenu. Les classes situées dans la partie Nord-Est de la fenêtre sont des classes présentes principalement dans les premières périodes, mais pas dans les dernières tranches de temps. A l'inverse, les classes situées dans le cadran Nord-Ouest laissent à penser qu'il s'agit de classes émergentes, qui sont en pleine extension, puisqu'elles n'étaient pas aussi développées dans les premières périodes. Les classes des zones Sud-Est et Sud-Ouest correspondent aux périodes de transition où les classes vont évoluer que ce soit au niveau des liens inter classes, mais aussi au niveau interne (évolution de la première période à la dernière). Les classes situées vers le centre de la figure sont les classes persistantes, présentes pour toutes les périodes. Alors que leurs liens inter classes varient au cours des quatre périodes, leur évolution interne au sein de chacune varie faiblement, comparé aux classes situées en périphérie.

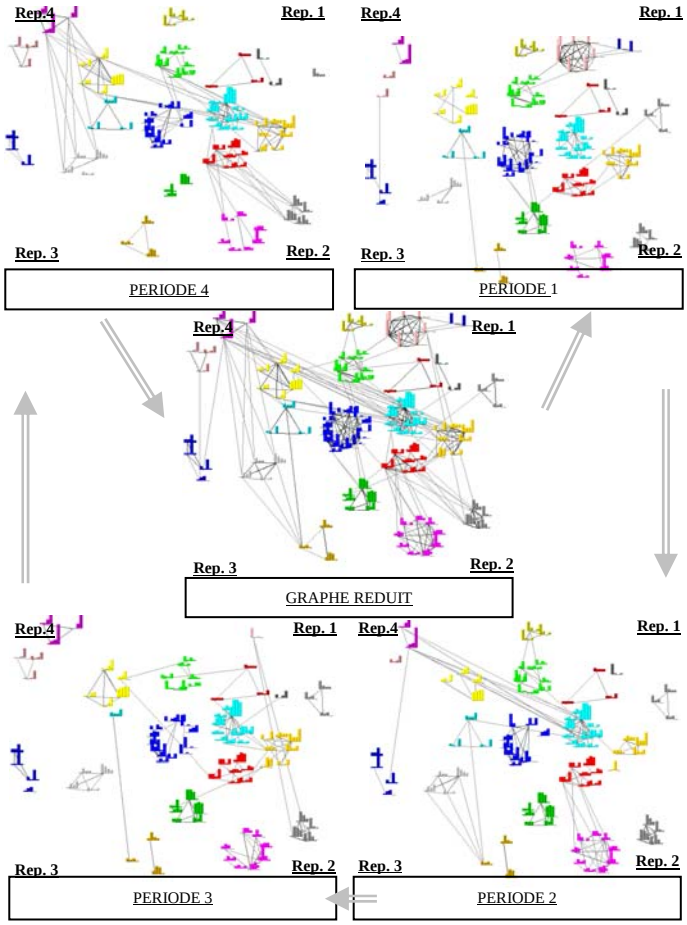

FIG. 1 : Visualisation du graphe global des classes, puis visualisation par animation des graphes de classe par période (« Rep n » ←→ repère temporel de la nième période).

5 Conclusion

Davantage qu'une simple recherche, la veille consiste à recueillir l'information, à la synthétiser et à tirer des conclusions aidant à la prise de décision en anticipant les tendances. L'outil VisuGraph permet le prétraitement des données, en particulier dans la gestion de la synonymie, et la synthèse des données relationnelles sous forme de matrices de cooccurrences décomposées en périodes homogènes. En se basant sur ces matrices, les données peuvent être représentées dans un premier temps, sous forme de graphe de classes global, toutes tranches de temps confondues, puis dans un second temps successivement sous forme de graphe de période. De par la prise en compte de la dimension temporelle, la position des sommets du graphe est stratégique et spécifiques aux caractéristiques liées au temps, telles que l'appartenance à certaines périodes et non à d'autres. Ainsi, chaque portion de la fenêtre de représentation correspond à une typologie relative au temps particulière. Cependant les limites de cette proposition concerne la méthode de classification choisie, le Markov Clustering MCL, qui ne prend pas en compte le point de vue de l'utilisateur, et, dans certains cas, le

partitionnement est soit trop fin soit trop grossier (à la limite une seule classe est trouvée). Il conviendrait d'offrir à l'utilisateur la possibilité d'intervenir dans la classification par une analyse visuelle des regroupements qui s'opère lorsqu'on joue sur les paramètres permettant de dessiner au mieux le graphe.

6 Références

Dousset B., Benjamaa T. (1988). Trilogie logiciel d'analyse de données. Conférence sur les systèmes d'informations élaborées.

Loubier E., Carbonnel S. (2007), Influence du prétraitement textuel sur la représentation graphique dans un contexte d'analyse de données relationnelles, VSST'07, Maroc, CD.

Eades P. (1984). A heuristic for Graph Drawing. Congressus Numerantium, vol. 42, p. 149-160.

Loubier E. (2007), La prise en compte de la dimension temporelle dans la visualisation de données par morphing de graphe, VSST'07, Maroc, CD.

Alpert C.J., Kahng A.B. (1995). Recent developments in netlist partitioning : A survey. The VLSI journal, vol. 19, pp.1-18.

Kuntz P., Henaux F. (2000). Numerical comparaison of two spectral decomposition for vertex clustering. Data Analysis, Classification and Related Methods, Proceeding Of IFCS'2000, Springer Verlag, pp.581-586.

Jouve B., Kuntz P. et Velin F. (2001). Extraction de structures macroscopiques dans des grands graphes par une approche spectrale. ECA, Hermès Science publication édition, vol. 1, pp. 173-184.

Karypis G., Kumar V. (1998). Multilevel k-way partitioning scheme for irregular graphs. Journal of Parrallel and distributed Computing, vol. 48, pp.96-129.

Van Dongen S. (2000). Graph Clustering by Flow Simulation. Thèse de doctorat, Université d'Utrecht, Allemagne.

Karouach S., Dousset B. (2003). Les graphes comme représentation synthétique et naturelle de l'information relationnelle de grandes tailles. Dans : Workshop sur la recherche d'information, associé à INFORSID'2003, Nancy, 03/06/2003-06/06/2003, INFORSID, p. 35-48, juin 2003.

Enright A.J., Van Dongen S. et Ouzounis C.A. (2002). An efficient algorithm for lagre-scale detection of protein families. Nucleic Acids Research, vol. 30, pp. 1575-1584.

Summary

Visualization of relational data in the form of optimized graph facilitates the analysis and the interpretation by the non-expert user. However, these graphs can easily become too complex to study globally. Then, it is necessary to divide them to facilitate the reading and the analysis. The solution is to simplify them, in a reduced graph whose tops represent each one a distinct group of tops: actors or terms of the studied field. In a second time, it is necessary to break up them into sub-graphs to take temporal dimension into account. The strategic watch platform Tetralogie makes it possible to synthesize the evolutionary relational data in the form of 3D co-occurrence matrices. VisuGraph, its visualization module is based on a Markov Clustering technique, which leads to a visualization in the form of reduced graph. We propose to take temporal dimension into account in our process of data classification. Thus, by the successive visualization of the various authorities, it becomes easier to analyze the intra (but also inter) classes evolution.