La visualisation de données relationnelles au service de la recherche d'informations

Eloïse Loubier— Wahiba Bahsoun

Institut de Recherche en Informatique de Toulouse
IRIT-SIG, Université Paul Sabatier,
118 route de Narbonne, 31062 Toulouse cedex 9

{loubier,wbahsoun}@irit.fr

Catégorie : chercheur

RÉSUMÉ. *Dans le cadre de la recherche d'informations, la restitution des documents s'effectue selon leur score de pertinence calculé, correspondant à une requête précise. Cependant des questions se posent quant à la représentation des documents et des requêtes ainsi que leur mise en correspondance. Le graphe est utilisé comme moyen de représentation et de visualisation de données, sans nécessiter de pré requis mathématique particulier. Cet article présente les apports de la visualisation d'information à la recherche d'information, en s'attachant au processus de conception d'un outil de visualisation et d'analyse développé par l'équipe SIG/EVI de l'IRIT le prototype VisuGraph. Dans ce contexte, nous proposons de nouvelles approches pour la visualisation des réseaux d'informations, afin d'en extraire avec facilité les documents les plus pertinents. Nous développons cette approche et nous illustrons notre contribution par des exemples concrets.*

ABSTRACT. *Within the framework of the information retrieval, the restitution of the documents is carried out according to their calculated relevance score. They correspond to a precise request. However questions arise as for the representation of documents and requests like their mapping. The graph is used as data visualization, without requiring mathematical knowledge. This article presents the contributions for the information visualization. We stick to the design process of a visualization and analysis tool developed in our laboratory IRIT by the SIG/EVI team: VisuGraph prototype. In this context, we propose new approaches for the information networks visualization, in order to extract easily the most relevant documents. We develop this approach and we illustrate our contribution by real examples.*

MOTS-CLES : *Recherche d'information, pertinence, dessin de graphe, visualisation des résultats de recherches, analyse relationnelle, réseaux sémantiques.*

KEYWORDS : *Information retrieval, relevance, graph drawing, visualisation of search results, relational analysis, semantic networks.*

1. Problématique

Pour être en mesure d'offrir aux utilisateurs les renseignements correspondants à leurs attentes, une solution de recherche d'information doit s'appuyer sur la pertinence des documents. Il est important de trouver la représentation la plus judicieuse pour la mise en correspondance des documents et des requêtes. La visualisation de données permet de représenter, révéler et analyser les structures sous jacentes entre les composants de la requête et les documents. La découverte d'informations pertinentes s'appuie sur les liens fonctionnels et sémantiques entre documents, acteurs, terminologie et concepts d'un domaine. Les relations entre ces derniers sont représentées par le biais de graphes aux dessins optimisés afin d'identifier plus facilement les topologies remarquables.

Cet article présente les apports de la représentation graphique de données à la recherche d'information, en s'attachant au processus de conception d'un prototype de visualisation et d'analyse.

Tout d'abord, nous exposons le concept de visualisation d'informations, ensuite le prototype VisuGraph développé par l'équipe SIG/EVI. Ce dernier est très puissant et permet la représentation graphique, ainsi que la classification interactive des données relationnelles. Le processus de visualisation peut être divisé en trois étapes : le prétraitement, la représentation et la navigation. Pour chacune de ces étapes, nous étudions les travaux effectués dans le domaine, mais, aussi, nous montrons les approches, mises en place dans VisuGraph, pour la visualisation des réseaux d'informations, afin d'en extraire avec facilité les éléments dominants, mais aussi les structures significatives des données.

Dans un second temps, nous expliquons notre contribution, qui consiste à mettre en pratique la représentation des documents et des mots-clés, dans un contexte de recherche d'information, afin de visualiser les différents regroupements de documents analysables. Nous présentons les techniques graphiques mise en place pour faciliter la détection d'éléments pertinents. Nous illustrons, par la suite, notre contribution, par des exemples concrets, appliqués au domaine de la recherche d'information.

Nous développons cette approche en mettant l'accent sur la structure des données, l'ergonomie de l'interface (visualisation, interactivité, point de vue), l'optimisation du dessin de graphe, de son animation.

2. La visualisation de données relationnelles

2.1. *Concept*

La représentation graphique permet de compléter la recherche d'information en visualisant une grande quantité d'informations de façon compréhensible et en fournissant au lecteur un maximum de renseignements synthétiques, qui ne sont que très rarement explicités dans les données brutes. La représentation graphique est un excellent vecteur d'analyse des données complexes, (Tufte, 1983), (Tufte, 1990), (Tufte, 1997).

Par exemple, on peut poser la question : Existe t'il des regroupements dans ce réseau ? La visualisation graphique peut nous donner une vue sur l'organisation des données ou en faire apparaître les propriétés structurelles pour la question tel élément est-il important dans le réseau ? Ces tâches d'analyse seraient très difficiles, voire impossibles, en basant l'analyse sur du texte brut, en particulier quand la taille de la donnée est importante.

Toute représentation visuelle de l'information, possède un certain degré d'interactivité, ainsi qu'une capacité à transmettre de l'information complexe à haute densité. Un graphe G = (V, E) est un ensemble de sommets V et d'arêtes E, joignant chaque paire de sommets. Ces derniers sont généralement représentés sous forme de cercles, reliés par des arcs sous forme de courbes ou segments. Il existe plusieurs types de représentations graphiques, suivant les objectifs de la visualisation. Principalement, le dessin d'un graphe s'effectue en suivant les règles établies par (Fruchterman et al., 1991), telles que :

– Disposition des sommets dans la fenêtre de représentation.

– Minimisation des croisements des arêtes.

– Respect d'une certaine symétrie dans la disposition du graphe (répartition équitable).

En se basant sur le dessin de graphe obtenu, les topologies remarquables sont identifiées, révélant les relations entre les différents acteurs (auteurs, laboratoires, entreprises, pays) et les termes et/ou les concepts d'un domaine.

Certains auteurs proposent toutefois l'idée qu'une correspondance existe entre notre représentation des similarités dans l'espace conceptuel et nos modalités de perception de l'environnement dans l'espace euclidien. Les entités conceptuelles possédant un taux d'attractivité élevé (liens forts ou haute similarité) auront tendance à se regrouper dans l'espace conceptuel et à se « tenir ensemble », (Gärdensfors, 2000).

Quoiqu'il en soit, les capacités de l'interface de visualisation à traduire et à faciliter l'interaction entre les informations à transmettre et l'observateur sont d'une importance capitale et ceci est d'autant plus critique lorsque les données à représenter sont textuelles. Cette spécificité de ces dernières exige que l'outil de visualisation, servant à représenter l'information contenue dans les réseaux sémantiques, possède des fonctionnalités efficaces d'aide à la manipulation des résultats.

2.2. Le prototype VisuGraph

2.2.1. *Principe*

Selon (Tufte, 1983), " Un excellent graphique est celui qui fournit au lecteur un nombre maximum d'idées dans le plus court lapse de temps en utilisant le moins d'encre et le plus petit espace possible ".

Basé sur ce principe, VisuGraph a été développé par (Karouach et al., 2004) et nous proposons d'en étendre les fonctionnalités de représentation, afin de l'adapter à la recherche d'information, par l'ajout d'un accès aux données textuelles, un algorithme de calcul de transitivité, des choix de visualisation prenant en compte les valeurs des métriques des données, développés dans les paragraphes suivants.

Les relations sont représentées à l'aide d'un graphe dont les sommets représentent les objets et les arêtes les liens assimilés à des ressorts. Le graphe est optimisé tout au long de sa représentation permettant le non chevauchement des sommets et croisement minimal des arêtes. La représentation des données s'effectue tout en laissant le choix à l'utilisateur, des fonctionnalités à appliquer sur le graphe telles que visualisation circulaire, affichage par seuil,…), pour obtenir davantage d'informations sur la structure des données, ciblées sur son axe de recherche comme la détection d'acteurs importants.

Il peut aussi choisir le mode de représentation des sommets, traduisant leur importance pour chaque période, sous forme de nuance, de cercle ou d'histogramme proportionnels à la métrique du sommet.

2.2.2. *Matrice de co-occurrences*

Les données relationnelles que nous traitons sont issues d'un processus de traitement d'information effectué sous la plateforme Tétralogie, mise en place par (Dousset *et al.*, 1988). Ces informations, issues de corpus, sont représentées sous formes matricielles par croisements d'entités, dont le contenu correspond aux co-occurrences. Ces matrices révèlent les groupes de mots apparaissant fréquemment ensemble. En général, on peut faire varier au moins un des constituants sur l'axe paradigmatique.

Considérons l'exemple d'une matrice, qui par croisement d'auteurs, indique le nombre d'articles co-écris par chaque paire. Nous avons alors une matrice de co-occurrence entre quatre auteurs {a, b, c, d}. Cette dernière indique le nombre d'articles, appelé nombre d'occurrences, co-écrits par deux auteurs. Ainsi le nombre d'occurrences entre les auteurs « a » et « d » pour la période 1 vaut 1. Notre démarche consiste alors à transformer ces données en une représentation sous forme de réseaux, dont les sommets représentent les auteurs {a, b, c, d} et les liens définissent les relations entre elles, comme le montre la figure 1.

	a	b	c	d
a	3	1	0	1
b	1	2	0	1
c	0	0	1	0
d	1	1	0	3

Tableau 1. *Matrice de cooccurrence*

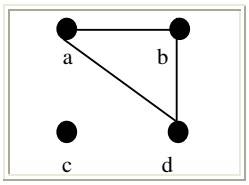

Figure1. *Représentation sous forme de graphe de la matrice de croisement des auteurs.*

2.2.3. *La représentation graphique*

Afin de placer au mieux les sommets, nous avons recours aux fonctions d'attraction et de répulsion des sommets.

Selon (Eades , 1984), un graphe est comparable à un modèle de ressort en s'inspirant des lois physiques pour dessiner le graphe. Un tel système engendre des forces entre les sommets, ce qui entraîne leur déplacement. En simulant des frottements qui diminuent progressivement l'énergie initiale, le graphe se stabilise sous une forme plus lisible. La force d'attraction entre les sommets peut être proportionnelle à la force du lien entre eux. Pour deux sommets υ_i et υ_j, elle est donnée par :

$$fa(\upsilon_i, \upsilon_j) = \beta_{ij} \times d_{ij} \alpha a / K \quad [1]$$

β_{ij} est fonction du poids de l'arête (υ_i, υ_j) et du poids des sommets υ_i et υ_j. Le facteur K est calculé en fonction de l'aire du dessin et du nombre de sommets du graphe et d_{ij} est la distance entre υ_i, υ_j dans le graphe. Si les sommets υ_i, υ_j ne sont pas reliés par une arête alors $fa(\upsilon_i, \upsilon_j) = 0$. La force de répulsion entre deux sommets υ_i et υ_j est définie par :

$$fr(\upsilon_i, \upsilon_j) = -K^2 / d_{ij} \alpha r \times \beta_{ij} \quad [2]$$

La variable αr, resp. αa, est une constante qui sert à définir le degré d'attraction (resp. répulsion) entre deux sommets.

Nos expérimentations nous ont permis d'attribuer des valeurs aux paramètres αr (codé « re » sur la figure 2), αa (codé « ra » sur la figure 2), βij (codés respectivement « ac », « rc » sur la figure 2) permettant la mise en place d'un graphe lisible, c'est-à-dire dont les croisements d'arêtes sont minimisés et la visibilité des clusters est maximale. A partir d'un état initial de forte énergie, nous appliquons ces forces jusqu'à ce que les sommets se positionnent harmonieusement les uns par rapport aux autres sans se superposer.

ae = 1.5
ac = 1.1
re = 2.3
rc = 2.2

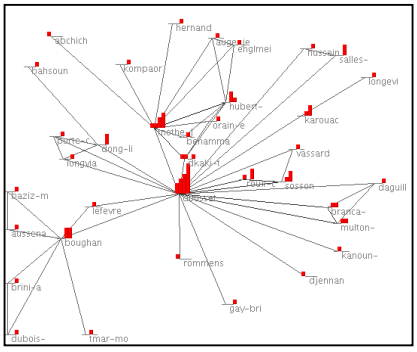

Figure 2. *Application des forces d attraction et de répulsion, dont les valeurs des coefficients ont été judicieusement choisies.*

2.2.4. *Partitionnement de graphe*

Suivant l'algorithme MCL (Markov Cluster algorithm) de Stijn van Dongen (van Dongen, 2000), pour des graphes de grande taille, VisuGraph propose un partitionnement de l'ensemble des sommets, ce qui conduit à travailler sur un graphe réduit plus facilement manipulable et plus clair. La limitation du nombre de sommets à afficher améliore la rapidité d'exécution. Chaque classe peut être visualisée soit séparément, soit avec ses connecteurs la liant aux autres classes, soit dans le contexte général reconstruit, à partir du graphe partiel, en figeant un représentant par classe.

La technique de partitionnement de Markov (MCL) repose sur une idée simple : un parcours aléatoire d'une partie dense d'un graphe a peu de risque de quitter cette partie dense avant d'avoir visité bon nombre de ses sommets. Plutôt que de simuler des marches aléatoires, l'algorithme propose d'étudier le flux du graphe en se basant sur un processus de Markov. La matrice de transition est successivement élevée à la puissance e (simulant e marches) puis normalisée.

L'algorithme proposé converge vers un point fixe ou vers un état récurrent. Les composantes connexes du graphe induit par la matrice finale sont les classes de la partition.

Soit M la matrice d'adjacence du graphe G = (V, E), e le facteur d'expansion et r le facteur d'inflation. Les différentes étapes de l'algorithme sont décrites ci-dessous :

- M est élevée à la puissance e.
- Chaque élément de la matrice (poids) est élevé à la puissance r.
- Chaque poids est ensuite divisé par le poids total de la ligne (normalisation).
- L'algorithme est réitéré tant qu'aucun point fixe ou état récurrent n'est atteint.

La figure 3 montre le partitionnement réalisé à partir d'un graphe de grande taille.

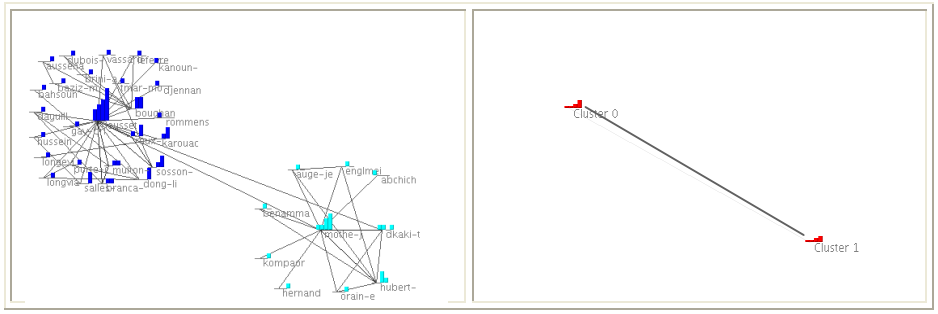

Figure 3. *Représentation des différentes classes puis visualisation partielle du graphe.*

2.2.5 *Fonctionnalités et paramétrisation*

L'ergonomie du prototype passe par la facilité d'accès aux différentes fonctionnalités, aussi bien dans leur sélection que dans leur application. Pour ce faire, nous avons réalisé un menu en deux parties distinctes.

- les fonctions permettent à l'utilisateur de spécifier, entre autres, les méthodes à appliquer à la représentation telles que le graphe circulaire, optimisé ou réduit, icônes représentant les sommets, coloration des sommets et des arêtes, affichage des noms des sommets, recherche d'un sommet spécifique, choix de la couleur de l'écran. La prise en compte de la transitivité sert à extraire des sous graphes qui permettent d'étudier l'environnement spécifique d'un sommet. Une fois celui-ci sélectionné, nous pouvons obtenir le nombre d'éléments connectés pour chaque degré de transitivité, la distance moyenne, la centralité, et afficher ou effacer, pas à pas les voisins.
- le paramétrage permet de contrôler le dessin avec la possibilité de modifier la profondeur de la transitivité, les coefficients des forces d'attraction et de répulsion, le seuil d'affichage, l'échelle des nuances des liens, la granularité du partitionnement.

La place occupée par le menu reste volontairement limitée, permettant un gain de place pour l'affichage du graphe.

3. La contribution

3.1. *Apport de la visualisation à la recherche d'information*

Comme nous l'avons vu dans le paragraphe 2.1, la représentation graphique permet de visualiser un nombre important de données et d'analyser leur structure. Les travaux de visualisation de (Marks *et al.*, 2005) s'appuient sur le constat suivant : un graphe peut représenter, de façon claire, plus de deux cents sommets alors qu'un écran d'ordinateur ne peut pas afficher plus de vingt lignes consécutives, résultantes d'un moteur de recherche classique.

Ainsi, il devient plus facile d'analyser les liens entre les sommets, ainsi que les différents regroupements de sommets. La visualisation globale des documents croisés aux mots-clés permet de révéler des informations non détectables dans les données brutes.

Nous nous sommes intéressés aux fonctionnalités qu'offrait l'outil DBL-Browser conçu par (Weber *et al.*, 2003), permettant la visualisation par auteur de sa bibliographie uniquement ou, séparément, le dessin de son graphe de collaboration avec d'autres auteurs, dans l'écriture de publications. Nous combinons ces deux modes de représentation en visualisant simultanément les auteurs mais aussi leurs actes, permettant ainsi d'observer directement l'ensemble de la bibliographie de l'individu, mais aussi ses alliances avec les autres auteurs.

Nous basons notre contribution selon le processus suivant :

- étape d'acquisition des données ;
- étape de diminution de la quantité des informations à manipuler, qui est très importante afin de ne conserver que la partie utile. Il est tout d'abord nécessaire d'extraire l'information sous forme de caractéristiques, par le biais de matrices de co-occurrences;
- étape de décision où l'on met en correspondance, par le dessin de graphe et à l'aide de règles convenablement choisies, les observations et les classes;
- étape d'évaluation des performances de la classification et de la pertinence des documents.

Appliquée à la recherche d'information, la visualisation des données par croisement d'informations peut être une aide précieuse, quant à l'évaluation de la pertinence des documents. La recherche d'information se base sur des requêtes, traduisant le besoin de l'utilisateur à travers des mots-clés. En effectuant un croisement entre les documents et le type des mots-clés (concepts, termes, noms d'auteurs..), nous pouvons détecter, via un graphe, les documents les plus proches et les plus liés aux différents composants de la requête, retrouvant avec aisance, les documents les plus pertinents.

Comme nous l'avons vu dans le paragraphe 2.2, les données en entrée sont des matrices de co-occurrences, qui peuvent être caractérisées comme indiquant simplement la présence/absence d'un lien, cas des matrices binaires, ou alors des matrices indiquant la valeur de la métrique du lien, c'est-à-dire une matrice d'occurrences, appelées aussi « matrice de comptage ».

Afin de faire ressortir graphiquement les différents regroupements de données, en vue de les analyser, nous avons recours à la méthode suivante. La "clusterisation" est une méthode statistique d'extraction de groupes (clusters) de termes ou d'expression de documents textuels. Cette méthode repose sur un calcul de fréquence d'apparition pour deux termes coexistants dans un même contexte (cooccurrence des termes). Les clusters ainsi formés sont significatifs car ils mettent en évidence les thématiques présentes dans les documents. La clusterisation génère ainsi des dizaines, voire plus, de clusters d'expressions liés les uns aux autres. Les relations entre les clusters sont ensuite mises en scène graphiquement : la cartographie sémantique. Dans la pratique de VisuGraph, la clusterisation est, accentuée par l'application d'une forte force d'attraction et d'une faible force de répulsion, impliquant une importante attirance des sommets liés. Plus le nombre d'occurrences entre les deux sommets est élevé, plus ils sont attirés. Ainsi, nous pouvons observer une certaine classification qui s'effectue par l'attirance des sommets les plus fortement liés. Nous obtenons alors un graphe organisé, pour lequel nous obtenons des classes ayant des caractéristiques communes.

3.2. *Détection de la pertinence*

3.2.1. *Codage de l'information*

Dans un graphe, tous les éléments (sommets et liens) n'ont pas la même importance ou le même rôle dans la structure locale ou globale du graphe. Pour pouvoir identifier visuellement les caractéristiques de chacun de ces éléments, il est indispensable d'introduire des variables visuelles afin de rendre la représentation du graphe plus riche en information. Pour cela, nous utilisons la couleur (ou intensité de couleur). Plus le lien entre deux sommets sera important, plus il sera représenté avec une forte intensité de couleur et inversement.

Cette fonctionnalité est réglable à travers un slider gradué, dont la valeur par défaut permet un dessin des liens non agressif pour la visualisation de l'utilisateur.

Ainsi la figure suivante illustre, à partir d'un extrait de matrice de co-occurrence, ce principe d'intensité de la couleur des liens.

	a	b	c	d
a	12	2	2	7
b	2	11	7	1
c	2	7	9	0
d	7	1	0	9

Tableau 2. *Matrice de cooccurrences*

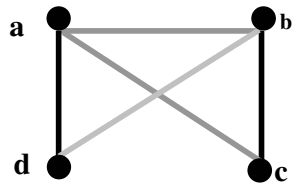

Figure 4. *Codage de la métrique des liens*

3.2.2. *Affichage des valeurs*

Dans l'objectif de visualiser davantage d'informations et de justifier la pertinence de tout document, nous ajoutons l'affichage optionnel de la valeur des liens entre sommets. Ainsi, lors de la représentation graphique d'une matrice de documents/termes, nous pouvons justifier la détection de documents pertinents par l'intensité des couleurs, vue dans le paragraphe précédent, mais aussi par la valorisation du lien. Nous obtenons directement le nombre d'occurrences du terme dans le document, permettant ainsi de définir si ce dernier est riche en information pour le terme choisi.

3.3. *Matrices de présence/absence*

Ces matrices sont binaires et indiquent simplement si des liaisons existent entre les données, sans valoriser l'importance quantitative de ces derniers. Si nous prenons l'exemple d'une matrice binaire croisant le champ « auteur » avec le champ « publication », nous obtenons la présence/absence des liens entre les différents actes et auteurs, bien représentés visuellement dans le graphe de la figure 5.

Nous visualisons si un individu a participé à l'écriture d'une publication ou non par la présence d'un lien entre un nom et un article. Ainsi sont révélés pour chaque article ses différents co-auteurs mais aussi les différentes collaborations de chaque auteur. Cette notion de relation entre co-auteur se base sur les travaux portant sur les réseaux sociaux étudiés par (Watts, 2004) et (Staab, 2005). Notre recherche est enrichie, puisqu'en visionnant notre graphe, nous pouvons découvrir l'ensemble des actes composés par l'individu choisi. Dans notre exemple, les auteurs sont représentés de couleur claire et les articles sont visualisés par une nuance plus foncée. L'application d'une forte force d'attraction mène à distinguer les différentes classes, constituées généralement de l'article et des co-auteurs. Deux classes peuvent être reliées par un auteur. En effet, si un individu a co-écrit l'article visualisé dans une première classe, mais aussi celui figurant dans la seconde, alors il se trouve comme étant la jointure entre les deux.

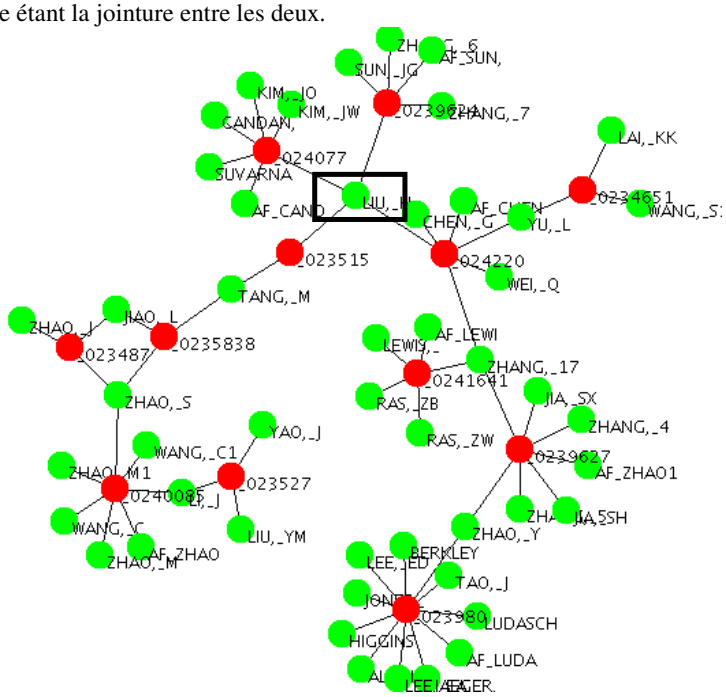

Figure 5. *Graphe des auteurs (sommets clairs) et de leurs publications (sommets foncés).*

Si nous regardons en particulier l'auteur LIU, encadré de noir dans la figure 6, nous pouvons voir qu'il a écrit quatre articles. Si nous étudions ses collaborations, nous constatons que Liu a collaboré avec plusieurs équipes, composées de nombreux auteurs. Il est relié à quatre classes, bien représenté par les quatre articles qui l'entourent.

Nous ajoutons à cette représentation une fonctionnalité, permettant, par sélection du document, de l'ouvrir dans un fichier texte afin de pouvoir lire son contenu. Ainsi, l'utilisateur est à même de pouvoir juger si le document sélectionné correspond bien à son besoin d'information. Si nous cliquons, sur le sommet représentant le document 024077 liant l'auteur Liu et Survana, nous pouvons visualiser l'acte sur la figure 6.

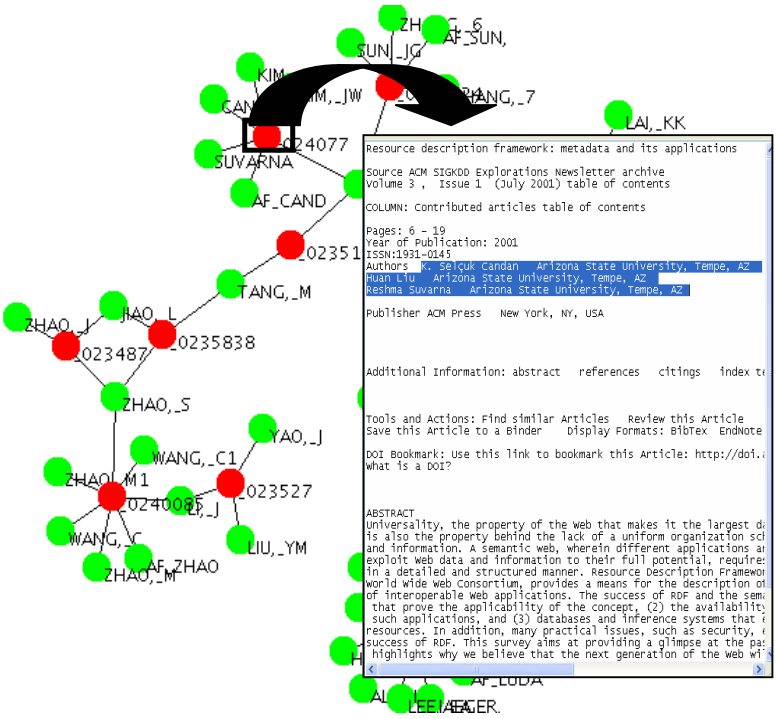

Figure 6. *Visualisation de l'article sélectionné (sommet foncé, encadré), écrit par Liu.*

3.4. *Matrices d'occurrences multiples*

3.4.1. *Détection de documents pertinents*

Nous nous intéressons aux matrices qui attribuent une valeur de métrique au lien entre deux champs, résultat d'un comptage. Si nous prenons le cas d'une matrice, croisant les champs « documents » et « mots-clés », comme dans la figure 7, les liens symbolisent le nombre d'occurrences du mot-clé choisi, dans les documents.

Ces derniers sont représentés par des cercles de couleur foncée alors que les mots-clés sont représentés par des sommets de couleur claire. Si nous ciblons le mot-clé de notre requête, nous pouvons voir directement le nombre mais aussi la pertinence des documents qui le contiennent à travers l'intensité du lien, signifiant le nombre d'occurrence de ce terme dans chacun d'eux, par, comme nous l'avons vu dans le paragraphe 3.2.1. Ainsi nous avons une restitution des résultats de notre requête sous forme de graphe.

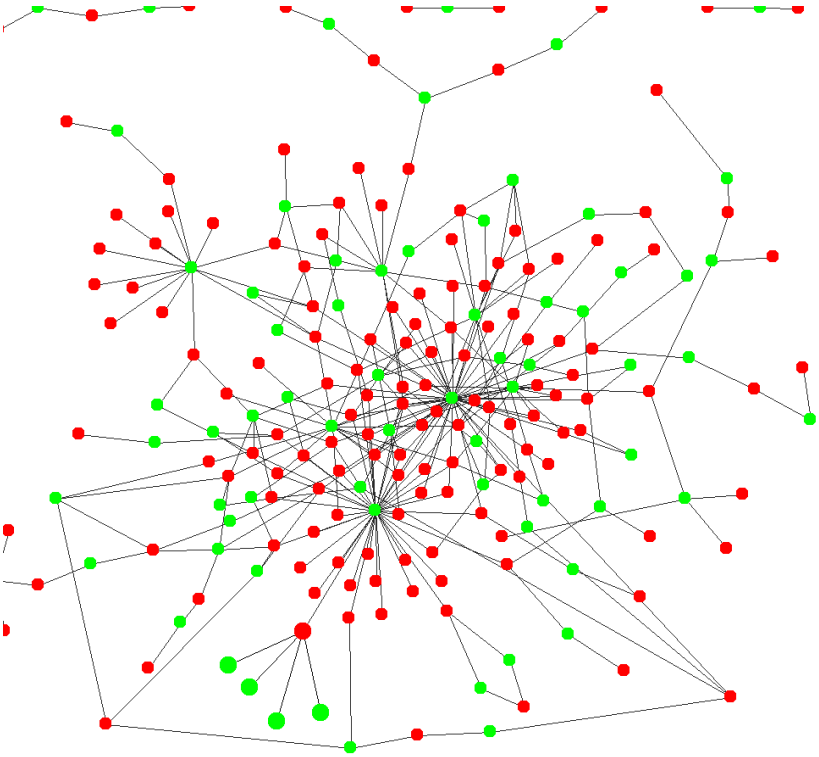

Figure 7. *Graphe des documents et des mots-clé, permettant pour chaque terme de retrouver tous les corpus les contenant individuellement.*

Si nous agrandissons une zone de ce graphe, en accentuant particulièrement l'intensité de coloration des liens selon leur valeur et que nous affichons le nom des documents, ainsi que des différents termes, nous obtenons le graphe 8.

Dans ce dernier, nous pouvons observer que le terme de la requête « neural » est relié par un lien de couleur foncée au document codé « 024175 ». Ses liens avec « 024225 » et « 0238412 » sont beaucoup plus clairs. Cela montre que le terme « neural » a été trouvé un plus grand nombre de fois dans le premier document, que nous jugeons plus pertinent que les deux autres.

De plus, l'affichage des valeurs des liens permet de confirmer cette affirmation. En effet, le lien entre « neural » et le document « 024175 », révèle que ce mot y est présent neuf fois alors qu'il n'apparaît que deux fois dans les deux autres. Le clic de souris sur le sommet « 024175 » permet l'affichage du document sous un éditeur de texte, surlignant de couleur claire les occurrences du terme « Neural » de la requête.

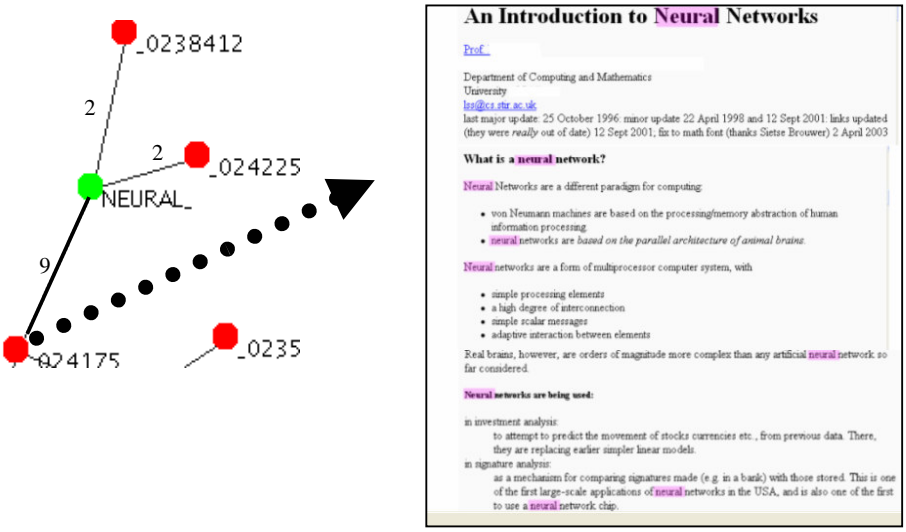

Figure 8. *Zoom du graphe de la figure 7, sur lequel nous affichons les noms des documents et les termes, et nous accentuons l'intensité des liens en affichant leur valeur.*

3.4.2. *L'expansion de requête*

Afin de préciser l'approche concernant l'expansion de requête, nous visualisons en figure 9, un extrait de la matrice représentée par le graphe de la figure 7. Nous avons appliqué au graphe 9 les forces des liens avec un coefficient de répulsion fort, afin de pouvoir isoler les termes des documents, en laissant apparaître la classification des données, comme nous l'avons vu précédemment. Nous avons effacé les valeurs des liens afin de rendre plus lisible la représentation.

En effet sur un tel graphe, l'affichage des noms des documents, mais aussi des termes et des valeurs des liens surchargent le dessin. Nous préférons ainsi juger de la pertinence des documents par rapport à l'intensité de couleur des liens, gagnant ainsi en lisibilité sur le graphe, en limitant le nombre d'affichages.

Graphiquement, l'expansion de requête se traduit par la transitivité. Dans un premier temps, nous sélectionnons un sommet, correspondant au terme de la requête. En augmentant, par le biais d'un slider réglable, le niveau de transitivité, nous extrayons le sous graphe, constitué du sommet sélectionné, mais aussi des éléments qui lui sont connectés, ainsi que leurs liens. L'affichage des voisins s'effectue alors pas à pas, visualisant progressivement tous les documents et les termes proches du mot-clé initialement choisi. Appliqué à la recherche d'information, la notion de transitivité se traduit par la détection du sous graphe composé d'un mot-clé de la requête et des documents qui sont reliés à ce dernier.

Plus nous augmentons la transitivité, plus nous ajoutons à ce sous graphe les mots-clés les plus proches de celui initialement choisi, ainsi que les documents qui lui sont associés. Nous illustrons ce principe à travers l'exemple suivant.

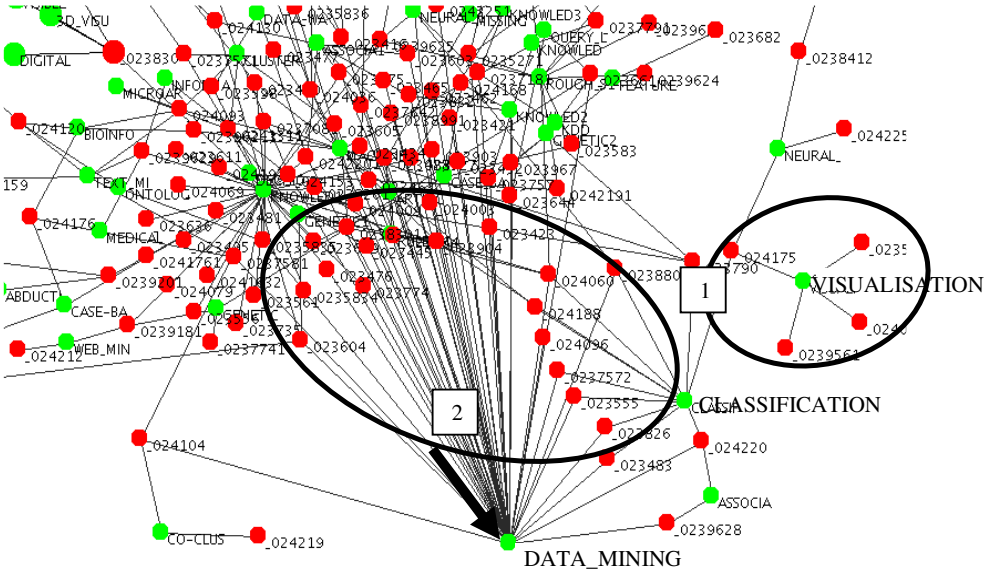

Figure 9. *Représentation de l'extrait de la matrice des documents (sommets foncés) et des mots-clés (sommets clairs).*

Dans cet exemple, nous allons considérer que l'utilisateur a effectué une requête, constituée du mot-clé « visualisation ». Nous constatons alors que quatre documents apparaissent, dans le cercle numéroté 1 sur la figure, comme pertinents, puisqu'ils sont liés directement à ce mot-clé. Si l'utilisateur effectue une requête

composée des mots-clés « visualisation » et « data_mining », nous observons que ces deux termes sont individuellement reliés à de nombreux documents. Or, plusieurs de ces derniers ne sont pertinents que pour un mot-clé, puisqu'ils ne sont pas reliés au second terme. Si nous prenons en compte la transitivité, pour passer de « visualisation » à « data_mining », nous voyons que le seul terme séparant ces deux mots-clés est « classification ». Ces trois mots clés étant liés, nous sommes en mesure de penser que leur association agrandit le domaine des réponses, enrichissant les connaissances restituées. Cela nous mène à réaliser une extension de requête, comme un traitement pour "élargir" le champ de recherche pour cette requête. Une requête étendue va contenir plus de termes reliés. Nous avons la chance de sélectionner plus de documents pertinents avec une requête étendue.

Dans notre figure, nous voyons clairement le premier groupe de documents, noté 2, qui ne sont liés qu'au terme « data_mining » ; le second groupe, noté 1, est directement lié à « visualisation ». Si nous rajoutons à notre requête le terme « classification », nous voyons que nous obtenons de nombreux documents, reliés à plusieurs mots-clés de la requête et comprenant les deux groupes obtenus précédemment. L'apport en information est alors plus riche et le domaine d'étude est davantage ciblé.

Tout comme pour l'exemple étudié dans le paragraphe précédent, nous pouvons cliquer sur un sommet correspondant à un document et obtenir le contenu, vérifiant ainsi la pertinence réelle de ce dernier.

4. Conclusion

Cet article présente une technique de visualisation de documents (articles, contenus textuelles) et de mots-clés (auteurs, concepts..) permettant la détection graphique des informations les plus pertinentes. Cette nouvelle approche permet aussi d'effectuer l'expansion de requêtes, au travers de la visualisation des liens entre les différents termes et les documents, par la transitivité basée sur le premier terme choisi. Alors que les modèles classiques de recherche d'information se basent sur des calculs, notre approche se base sur l'analyse de graphe et sur la détection visuelle de documents pertinents.

De nombreuses spécificités apparaissent à la suite de ce travail. En effet, le constat de nos travaux tend à montrer qu'il est plus intéressant de travailler sur des matrices de « comptage » plutôt que sur des matrices binaires. En effet, les premières permettent de mettre l'accent sur l'importance du lien entre deux sommets, révélant ainsi la pertinence réelle d'un document pour un mot-clé choisi, contrairement aux secondes qui facilitent la symbolisation de présence/absence de lien entre les données.

VisuGraph apparaît comme un outil d'analyse des données, puissant et ergonomique et qui permet de révéler, comprendre et anticiper les structures sous-jacentes afin d'identifier leurs implications stratégiques et leur pertinence. Cependant, ce prototype pourrait être amélioré, puisqu'il ne prend pas en compte le contexte utilisateur. Ainsi, nous devrions situer précisément le profil utilisateur, afin de mettre en relief ses centres d'intérêts par une visualisation accentuée des structures qu'il privilégie.

5. Bibliographie

Dousset B., Benjamaa T., « Tétralogie logiciel d'analyse de données », *Conférence sur les systèmes d'informations élaborées : Bibliométrie – Information Stratégique – Veille technologique,* 1988.

Eades P., « A heuristic for Graph Drawing ». *Congressus Numerantium*, vol. 42, p. 149-160, 1984.

Fruchterman T., Reingold E., « Graph Drawing by Force-directed Placement », *SOFTWARE—PRACTICE AND EXPERIENCE,* Department of Computer Science, University of Illinois at Urbana-Champaign, 1304 W., VOL. 21(1 1), p. 1129-1164, 1991.

Gärdensfors P., « Conceptual spaces: The geometry of thought ». Cambridge, *MA : MIT Press*, 307 p, 2000.

Karouach S., Dousset B., « Analyse d'information relationnelle par des graphes interactifs de grandes tailles ». *EGC'04*, Clermont Ferrand, 2004.

Marks L., Hussell J., McMahon T., Luce E., « ActiveGraph : A digital library visualization tool », *Research Library*, Los Alamos National Laboratory, USA, Springer-Verlag, 2005.

Staab S., « Social networks applied ». *IEEE* Intell. Syst. 20(1), 80-93, 2005.

Tufte E., « The Visual Display of Quantitative Information ». *Graphics Press*, 1983.

Tufte E., « Envisioning Information ». Graphics Press, 1990.

Tufte E., « Visual Explanations ». *Graphics Press*, 1997.

Van Dongen S., « Performance criteria for graph clustering and Markov cluster experiments », *National Research Institute for Mathematics and Computer Science in the Netherlands*, Amsterdam 2000.

Watts D.J., « Six degrees: The Science of a Connected Age ». NY: W. W. Norton, 2004.

Weber A., Agarwal S., Fankhauser P., Gonzalez-Ollala J., Hartmann J., Hollfelder S., Jameson A., Klink S., Lehti P., Ley M., Rabbidge E., Schwarzkopf E., Shrestha N., Stojanovic N., Studer R., Stumme G., Walter B., « Semantic Methods and Tools for Information Portals », *GI Jahrestagung* (1), 116-131, DBLP:conf/gi/2003-1, http://dblp.uni-trier.de, 2003.