Interrogation de bases de données de graphes : Une approche basée sur un skyline par similarité

Katia Abbaci, Allel Hadjali, Ludovic Liétard, Daniel Rocacher

Interrogation de bases de données de graphes : Une approche basée sur un skyline par similarité

Katia Abbaci*, Allel Hadjali*
Ludovic Liétard**, Daniel Rocacher*

*IRISA/ENSSAT
Rue de Kérampont BP 80518 Lannion, France
Katia.Abbaci@enssat.fr
Allel.Hadjali@enssat.fr
Daniel.Rocacher@enssat.fr
**IRISA/IUT
Rue Edouard Branly BP 30219 Lannion, France
Ludovic.Lietard@univ-rennes1.fr

Résumé. L'un des problèmes fondamentaux des bases de données de graphes est la recherche de graphes similaires à une requête à graphe. Les approches existantes traitant ce problème s'appuient, généralement, sur une seule mesure de similarité entre les structures de graphes. Dans cet article, nous proposons une approche où la similarité entre graphes n'est plus un scalaire unique mais un vecteur de scalaires. Pour cela, nous utilisons le concept de *skyline par similarité* d'une requête à graphe défini par un sous-ensemble de graphes, de la base de données interrogée, qui sont *les plus similaires* à la requête au *sens de Pareto*. L'idée est d'effectuer une comparaison multidimensionnelle entre graphes en termes de d mesures de similarité locales et d'identifier les graphes qui sont *maximalement similaires* au sens d'une relation de dominance par similarité. Une méthode pour raffiner le résultat de la recherche est aussi discutée en s'appuyant sur le critère de *diversité entre les graphes*.

1 Introduction

Les requêtes dites skyline (Borzsonyi et al., 2001), ou tout simplement les requêtes skyline, visent à aider l'utilisateur à prendre des décisions intelligentes en présence de données complexes, en identifiant un ensemble de points intéressants où les critères pris en compte sont multiples, différents et souvent contradictoires. Soit un ensemble r d'objets ou de points multidimensionnels, une requête skyline retourne l'ensemble des points *non-dominés* dans r. Un point p domine un point q si p est au moins aussi bon que q pour toutes les dimensions et strictement meilleur que q dans au moins une dimension. Plusieurs recherches ont été réalisées pour le développement d'algorithmes efficaces et pour introduire différentes variantes de requêtes skyline (Khalefa et al., 2008; Pei et al., 2007; Yiu et Mamoulis, 2007; Hadjali et al., 2010). Pour autant que nous le sachions, peu de travaux sur les requêtes skyline existent dans le domaine des Bases de Données de Graphes (BDGs).

D'une autre part, les graphes sont devenus de plus en plus importants dans la modélisation des données structurées complexes dans plusieurs applications réelles et récents (Bioinformatique, Reconnaissance de formes, XML, Chimie, etc). Toutes ces applications indiquent l'importance et la large utilisation du paradigme des BDGs. D'une manière générale, on peut classer les requêtes adressées à une BDG en deux catégories : (1) *Recherche de graphes par relation d'inclusion* et (2) *Recherche de graphes par similarité*. La première catégorie se compose des deux sous-problèmes suivants : (i) *recherche de sous-graphes* : soit une BDG $D = \{g_1, g_2, \ldots, g_n\}$ et une requête à graphe *q* (dite requête sous-graphe), il s'agit de rechercher tous les graphes $g_i \in D$ tel que *q* est un sous-graphe de g_i (i.e., $q \subseteq g_i$) ; (ii) *recherche de super-graphes* : soit une BDG $D = \{g_1, g_2, \ldots, g_n\}$ et une requête à graphe *q* (dite requête super-graphe), il s'agit de rechercher tous les graphes $g_i \in D$ tel que *q* est un super-graphe de g_i (i.e., $q \supseteq g_i$). Les deux sous-problèmes font appel à la procédure de vérification d'isomorphisme de sous-graphes, qui est NP-Complet. Ainsi, plusieurs approches de traitement de requêtes à graphes utilisant des techniques d'indexation ont été développées pour réduire l'espace de recherche et résoudre efficacement ces deux sous-problèmes (Chen et al., 2007; Yan et al., 2004; Zhang et al., 2007, 2009).

Quant à la deuxième catégorie (i.e., recherche de graphes par similarité), qui consiste à rechercher tous les graphes d'une BDG qui sont structurellement similaires au graphe de la requête considérée, est apparue comme une nouvelle tendance pour les raisons suivantes (Tian et Patel, 2008; Petrakis et Faloutsos, 1997) : (i) de nombreuses et réelles BDGs sont de nature bruitée et incomplète, d'où la nécessité de *l'appariement*[1] *approximatif* de graphes ; (ii) plusieurs applications modernes préfèrent les résultats d'un appariement approximatif plutôt que ceux d'un appariement exact car ils fournissent davantage d'informations, comme ce qui pourrait être manquant ou superflu dans un graphe de requête ou dans une BDG.

Durant ces dernière années, un certain nombre d'approches ont été proposées pour répondre aux requêtes de recherche de graphes par similarité (ou simplement, requêtes par similarité). À titre d'exemple, *Grafil* (Yan et al., 2005), *C-Tree* (He et Singh, 2006) et *Tale* (Tian et Patel, 2008) sont trois techniques proposées pour répondre aux requêtes sous-graphes au moyen d'un appariement approximatif. Shang et al. (2010) ont proposé une technique pour répondre aux requêtes super-graphes en utilisant une recherche par similarité. Les deux approches *C-Tree* et *Tale* utilisent *la distance d'édition* pour mesurer la similarité entre graphes, tandis que les travaux réalisés dans (Yan et al., 2005) et (Shang et al., 2010) utilisent la notion de *sous-graphe commun maximal* pour le calcul de telle similarité.

Comme on peut le constater, les approches proposées et dédiées aux requêtes par similarité s'appuient uniquement sur une seule mesure pour évaluer la similarité entre graphes. Toutefois, un graphe est une structure complexe de nature et comprend une multitude de caractéristiques de base. Il est alors difficile de donner une définition significative de la similarité entre graphes en utilisant un seul scalaire. Dans cet article, nous préconisons que plusieurs indices sont nécessaires pour que la similarité entre graphes soit évaluée d'une manière significative et efficace. Chaque indice est dédié à mesurer une distance (ou similarité) locale entre deux graphes afférent à un aspect donné dans la structure de graphe. Ainsi, la similarité entre graphes est, maintenant, caractérisée par un vecteur de mesures de distance locales au lieu

1. Appariement de graphes correspond au processus spécifique de l'évaluation de la similarité structurelle entre deux graphes.

d'une seule mesure. De cette façon, on peut préserver l'information concernant la similarité sur chaque caractéristique lors de la comparaison de deux graphes.

Nous proposons une approche pour traiter les requêtes par similarité en utilisant le paradigme de skyline par similarité, issu du domaine de raisonnement par cas et proposé dans (Hüllermeier et al., 2008). D'une manière générale, *le skyline par similarité* d'une requête à graphe est défini par le sous-ensemble des graphes de la BDG interrogée qui sont *les plus similaires* à la requête au sens de Pareto. L'idée est d'effectuer une comparaison multidimensionnelle entre graphes en terme de *d* mesures de distance locales et d'identifier les graphes qui sont *maximalement similaires* au sens d'une relation de dominance par similarité.

L'article est structuré comme suit. La section 2 introduit quelques notions préliminaires. La section 3 aborde l'évaluation de similarité entre graphes. Dans la section 4, nous discutons la relation de dominance par similarité pour définir le skyline dédié aux requêtes à graphes. La section 5 présente un exemple détaillé. La section 6 propose une méthode pour raffiner les résultats retournés dans le skyline. La section 7 conclut l'article.

2 Notions préliminaires

2.1 Les requêtes skyline

Les requêtes Skyline (Borzsonyi et al., 2001) représentent un exemple spécifique et représentatif des requêtes à préférences. Elles s'appuient sur le principe de dominance de Pareto qui peut être défini comme suit :

Définition 1. Soit *r* un ensemble de points multidimensionnels et $p = (p_1, p_2, \ldots, p_d)$ et $q = (q_1, q_2, \ldots, q_d)$ deux points de *r*. On dit que *p domine* (au sens de Pareto) *q* ssi sur chaque dimension $p_i \leq q_i$ (pour $1 \leq i \leq d$) et sur au moins une dimension $p_j < q_j$.

Par souci de simplicité, et sans perte de généralité, nous supposons que plus la valeur p_i est petite, meilleure elle est. On dit alors que *p est préféré à q* et on dénote par $p \succ q$.

Définition 2. Le skyline de *r* est l'ensemble des points dominés par aucun autre point.

Les requêtes skyline calculent donc l'ensemble des tuples optimaux au sens de Pareto dans une relation, i.e., les tuples qui ne sont dominés par aucun autre tuple dans la même relation.

Exemple 1. Considérons une base de données contenant des informations sur des hôtels comme indiqué dans le tableau 1 (où la dimension $d = 2$).

Hôtel	Prix(€)	Distance (Km)
H_1	4.0	150
H_2	3.0	110
H_3	2.5	240
H_4	2.0	180
H_5	1.7	270
H_6	1.0	195
H_7	1.2	210

TAB. 1 – *Exemple d'hôtels.*

Considérons une personne qui cherche un hôtel aussi proche que possible de la plage et ayant un prix faible. On peut vérifier que le skyline résultant S contient les hôtels H_2, H_4 et H_6. Par exemple, H_1 est dominé par H_2, et H_7 par H_6.

2.2 Quelques définitions

Définition 3 (Graphe). Un graphe g est défini par un quadruplet *(V, E, L, l)* où V est l'ensemble des nœuds, E est l'ensemble des arêtes, L est l'ensemble des étiquettes et l est la fonction d'étiquetage qui met en correspondance chaque nœud ou arête avec une étiquette de L.

Pour simplifier la présentation, dans cet article, les graphes sont étiquetés et non-orientés. Notons que différents nœuds peuvent avoir la même étiquette et la taille de g est définie comme suit : |g| = |E(g)| (i.e., la taille d'un graphe est le nombre de ses arêtes).

Définition 4 (Isomorphisme de graphe). Soit deux graphes *g = (V, E, L, l)* et *g' = (V', E', L', l')*, *g* est *isomorphe à g'* (dénoté par $g \approx g'$) s'il existe une bijection $f : V \rightarrow V'$, telle que

1. $\forall$ *v* $\in$ *V, f(v)* $\in$ *V' et l(v) = l'(f(v))* ;
2. $\forall$ *(u, v)* $\in$ *E, (f(u), f(v))* $\in$ *E', et l(u, v) = l'(f(u), f(v))*.

Définition 5 (Isomorphisme de sous-graphe). Soit deux graphes *g = (V, E, L, l)* et *g' = (V', E', L', l')*, *g* est *isomorphe de sous-graphe à g'* s'il existe une injection $f : V \rightarrow V'$, telle que

1. $\forall$ *v* $\in$ *V, f(v)* $\in$ *V' et l(v) = l'(f(v))* ;
2. $\forall$ *(u, v)* $\in$ *E, (f(u), f(v))* $\in$ *E' et l(u, v) = l'(f(u), f(v))*.

Définition 6 (Sous-graphe v.s. super-graphe). Soit deux graphe *g = (V, E, L, l)* et *g' = (V', E', L', l')*, *g* est dit *sous-graphe de g'* (ou *g'* est un *super-graphe de g*), dénoté par $g \subseteq g'$ (ou $g' \supseteq g$), s'il existe un isomorphisme de sous-graphe de *g* à *g'*.

Définition 7 (Sous-graphe Commun Maximal, SCM). Soit deux graphes g_1 et g_2, le sous-graphe commun maximal de g_1 et g_2 est le plus grand sous-graphe connecté de g_1 qui est isomorphe de sous-graphe à g_2, dénoté par *g'= SCM(g_1, g_2)*.

3 Sur la similarité entre graphes

Plusieurs modèles ont été proposés (Bunke, 1997; Bunke et Shearer, 1998; Wallis et al., 2001) pour mesurer la similarité entre deux graphes. Chaque modèle est pertinent pour une classe d'applications, car il n'existe pas de modèle "*universel*" qui répond aux différents besoins de toutes les applications du monde réel. Ci-après, nous présentons, d'une part, les mesures [2] les plus utilisées pour déterminer les similarités entre graphes et, d'autre part, nous donnons l'interprétation de chaque mesure de similarité dans le cadre d'interrogation de BDGs.

3.1 La distance d'édition de graphes

La distance d'édition de graphes (Bunke, 1997) recherche l'ensemble d'opérations d'édition (insertion, suppression ou ré-étiquetage de nœuds et d'arêtes) nécessaires pour transformer, avec un *coût minimal*, un graphe en un autre graphe. Chaque opération est associée à une fonction de coût qui varie selon la quantité de distorsion introduite par la transformation.

2. Pour des raisons d'espace, l'analyse de la complexité du calcul de chaque mesure n'est pas abordée ici.

Définition 8 (Distance d'édition sous la fonction de mise en correspondance f) (Shang et al., 2010). La distance d'édition entre deux graphes *g=(V, E, L, l)* et *g'=(V', E', L', l')* sous la fonction de mise en correspondance *f : g* $\rightarrow$ *g'* est le coût de la transformation de *g* en *g'* :

$$d_f(g, g') = \sum_{u \in V} d(u, f(u)) + \sum_{e \in E} d(e, f(e)),$$

où *d(u, f(u))* (resp. *d(e, f(e))*) est la mesure de distance de nœuds (resp. d'arêtes) (par exemple, *d(u, f (u))* peut être considéré comme étant le *coût* résultant de la transformation de *u* en *f(u)*).

Les mesures de distance de nœuds et d'arêtes dépendent des domaines d'applications. En pratique, le choix des opérations d'édition élémentaires et leur coût représentent une tâche très difficile. Par souci de simplicité, nous considérons, dans cet article, une mesure de distance *uniforme* : la distance entre deux nœuds ou deux arêtes est égale à *1* si leurs étiquettes sont différentes ; *0* sinon.

Ainsi, la distance d'édition entre deux graphes peut être définie de la manière suivante.

Définition 9 (Distance d'édition de graphes). La distance d'édition entre deux graphes g_1 et g_2 est *la distance d'édition minimale* sous toutes les mises en correspondance possibles :

$$Dist_{Ed}(g_1, g_2) = min_f\{d_f(g_1, g_2)\} \quad (1)$$

Plus $Dist_{Ed}(g_1, g_2)$ est petite, plus les deux graphes sont similaires. On peut facilement vérifier que la distance d'édition entre deux graphes *isomorphes* est égale à *zéro*. Notons que ce type de distance ne souffre d'aucune restriction et peut être appliqué à tout type de graphes.

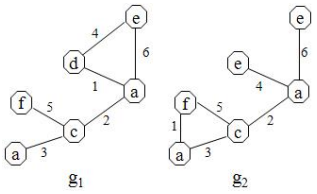

FIG. 1 – *Exemple de graphes étiquetés*

Exemple 2. Considérons les graphes étiquetés de la figure 1. La séquence d'opérations d'édition nécessaire pour transformer g_1 en g_2 est : (i) une suppression d'arêtes, i.e., l'arête *(d, e)*, (ii) un ré-étiquetage d'arêtes, i.e., changer l'étiquette de l'arête *(a, d)* de *1* à *4*, (iii) un ré-étiquetage de nœuds, i.e., changer l'étiquette du nœud *d* de *d* à *e*, (iv) une insertion d'arêtes i.e., l'arête *(a, f)* avec l'étiquette *1*. En utilisant des mesures de distance uniformes, on peut vérifier que cette séquence est la meilleure (i.e., la minimale). Ainsi, $Dist_{Ed}(g_1, g_2)$ = *4*.

Il est important de noter que, dans un contexte d'interrogation de BDGs, cette mesure de distance nous renseigne sur les caractéristiques qu'un graphe cible de la BDG et le graphe d'une requête, ne partagent pas.

3.2 La distance basée sur la notion de SCM

Bunke et Shearer (1998) ont développé un autre type de mesures de similarité entre graphes qui est basé sur *le sous-graphe commun maximal (SCM)*.

Définition 10 (Similarité basée sur le SCM). Soit deux graphes g_1 et g_2, la similarité basée sur le *SCM* est définie comme suit,

$$Sim_{SCM}(g_1, g_2) = \frac{|SCM(g_1,g_2)|}{|max(g_1,g_2)|},$$

où |*max*(g_1, g_2)|=*max*(|g_1|, |g_2|) et |*SCM*(g_1, g_2)| dénote le nombre d'arêtes dans *SCM*(g_1, g_2).

Clairement, plus le *SCM* de deux graphes est large, plus leur similarité est élevée. La mesure Sim_{SCM} est normalisée (i.e., $0 \leq Sim_{SCM}(g_1, g_2) \leq 1$) car |*SCM*($g_1$, g_2)| $\leq$ |*max*(g_1, g_2)|. Ainsi, la mesure de distance de graphes, $Dist_{SCM}$, dérivée de Sim_{SCM} s'écrit :

$$Dist_{SCM}(g_1, g_2) = 1 - Sim_{SCM}(g_1, g_2) \quad (2)$$

Il a été montré dans (Bunke et Shearer, 1998) qu'une telle mesure est une métrique et conduit à une distance ayant des valeurs dans l'intervalle [0, 1]. L'avantage principal de l'approche basée sur le *SCM* est la non utilisation de fonctions de coût, palliant ainsi l'inconvénient principal de l'approche basée sur la distance d'édition.

Exemple 3. Reprenons l'Exemple 2. Il est facile d'identifier le *SCM* (g_1, g_2), voir la figure 2. Par application de (2), nous obtenons (où |SCM(g_1, g_2)| = 4 et |max(g_1, g_2)|= 6) :

$Dist_{SCM}(g_1, g_2)$ = 1 - $\frac{|SCM(g_1,g_2)|}{|max(g_1,g_2)|}$=0.33,

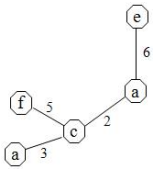

FIG. 2 – *SCM de g_1 et g_2.*

Dans un contexte d'interrogation de BDGs, le *SCM* de deux graphes véhicule de l'information sur les caractéristiques partagées par le graphe de la base interrogée et le graphe de la requête.

3.3 La distance basée sur l'UG

La mesure de distance basée sur l'union de graphes (UG), proposée par Wallis et al. (2001), est basée sur le principe de l'union de graphes. L'union de graphes (plutôt que le plus grand des deux graphes) est utilisée pour modéliser la taille du problème.

Définition 11 (Similarité basée sur l'UG). Soit deux graphes g_1 et g_2, la similarité de graphes basée sur l'union de graphes est définie comme suit,

$$Sim_{UG}(g_1, g_2) = \frac{|SCM(g_1,g_2)|}{|g_1|+|g_2|-|SCM(g_1,g_2)|},$$

où le dénominateur représente la taille de l'union des deux graphes selon une vue ensembliste [3].

3. Cette mesure de similarité ressemble à l'indice de Jaccard utilisé pour mesurer la similarité entre deux ensembles A et B, i.e., J(A, B) = |A ∩ B| / |A ∪ B|.

Cette mesure de similarité est aussi normalisée et son comportement est assez proche de celui de Sim_{SCM}. Il est facile de voir que $Sim_{UG}(g_1, g_2) \leq Sim_{SCM}(g_1, g_2)$ (ce qui signifie que Sim_{UG} est une mesure plus exigeante que Sim_{SCM}). L'utilisation de l'union de graphes est motivée par le fait que les changements dans la taille du *plus petit graphe* qui préservent le *SCM*(g_1, g_2) constant ne sont pas pris en compte dans $Sim_{SCM}(g_1, g_2)$, tandis que la mesure $Sim_{UG}(g_1, g_2)$ prend en compte cette variante.

La mesure de distance de graphes dérivée de Sim_{UG} s'écrit aussi sous la forme :

$$Dist_{UG}(g_1, g_2) = 1 - Sim_{UG}(g_1, g_2) \tag{3}$$

Il a été aussi prouvé que $Dist_{UG}(g_1, g_2)$ est une métrique et ayant des valeurs dans l'intervalle [0, 1].

Exemple 4. Reprenons encore les graphes de l'Exemple 2. Par application de (3), la distance basée sur l'*UG* entre g_1 et g_2 est

$Dist_{UG}(g_1, g_2)$ = 1 - $\frac{|SCM(g_1,g_2)|}{|g_1|+|g_2|-|SCM(g_1,g_2)|}$=0.33,

où |SCM (g_1, g_2)|= 4 (voir Exemple 3) et |g_1| = |g_2|= 6.

Dans un cadre d'interrogation des BDGs, ce type de similarité donne également des informations sur le nombre de caractéristiques communes entre un graphe de la base interrogée et le graphe d'une requête.

4 Skyline de graphes par similarité

Cette section est consacrée à la définition d'un concept appelé *skyline par similarité* pour supporter la recherche de graphes sans spécifier une mesure de similarité globale entre les structures de graphes.

Dans ce qui suit, nous supposons que la similarité entre graphes est une notion composée, i.e., un vecteur de mesures de distance. Chaque mesure peut être considérée comme une similarité locale exprimant la ressemblance des structures de graphes vis-à-vis d'une certaine caractéristique.

Définition 12 (Similarité composée de graphes, SCG). Soit *g* et *g'* deux graphes, une similarité composée entre *g* et *g'* est un vecteur de mesures de distance locales, dénotée par

$$SCG(g, g') = (Dist_1(g, g'), Dist_2(g, g'), \ldots, Dist_d(g, g')) \tag{4}$$

où $Dist_i$(g, g'), pour $i = 1, \ldots, d$, représente une mesure de distance locale entre graphes.

Soit $D = \{g_1, g_2, \ldots, g_n\}$ une BDG et *q* une requête par similarité (i.e., ce qui signifie que l'utilisateur est intéressé par les graphes de *D* qui sont les plus similaires à *q*). Comme il n'existe pas de mesure de similarité globale entre graphes, l'idée est de procéder à une comparaison multidimensionnelle entre graphes en termes de *d* mesures de distance (locales) pour rechercher les graphes qui sont maximalement similaires au sens de la relation de dominance par similarité définie par :

Définition 13 (Relation de dominance par similarité). Soit une requête à graphe *q* et deux graphes *g* et *g'*, on dit que *g' est dominé par similarité* par *g* dans le contexte de *q*, dénoté par $g \succ_q g'$, ssi les deux conditions suivantes sont vérifiées :

1. $\forall\, i \in \{1, \ldots, d\}$, $Dist_i(g, q) \leq Dist_i(g', q)$,

2. $\exists\, k \in \{1, \ldots, d\}$, $Dist_k(g, q) < Dist_k(g', q)$.

La relation $g \succ_q g'$ est vérifiée si g n'est pas moins similaire à q que g' dans toutes les dimensions et (strictement) plus similaire à q que g' dans au moins une dimension. On peut observer que g est potentiellement plus intéressant que g' comme graphe réponse. Par conséquent, l'ensemble des graphes les plus similaires à q sont ceux qui ne sont pas dominés (au sens de Définition 13). De tels graphes, appelés *graphes optimaux au sens de Pareto*, représentent ce que nous dénotons par *le skyline de graphes par similarité (SGS)* :

$$SGS(D, q) = \{g \in D |\ \nexists g' \in D, g' \succ_q g\} \tag{5}$$

où $g' \succ_q g$ signifie que g est dominé par similarité par g'.

Pour illustrer cette approche, nous présentons dans la section suivante un exemple où $d = 3$. $SCG(g, q)$ est alors un vecteur de trois composantes exprimées en termes des mesures de distance locales décrites dans la section 3, i.e.,

$$SCG(g, q) = (Dist_{Ed}(g, q), Dist_{SCM}(g, q), Dist_{UG}(g, q)).$$

5 Un exemple illustratif

Soit $D = \{g_1, g_2, g_3, g_4, g_5, g_6, g_7\}$ une BDG et q une requête par similarité, comme le montre la figure 3. Afin de retourner les réponses les plus intéressantes par rapport à q, on calcule le skyline de graphes par similarité *SGS(D, q)*. Voir les tableaux 2 et 3 qui résument les valeurs de $|SCM(g_i, q)|$ et les vecteurs de similarité de graphes $SCG(g_i, q)$, pour $i = 1, \ldots, 7$, respectivement.

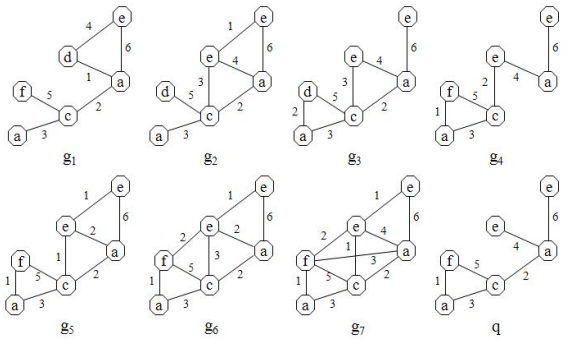

FIG. 3 – *La base de données D et le graphe requête q.*

Par application de (4), l'ensemble des graphes optimaux au sens de Pareto, i.e. le skyline de graphes par similarité, est donné par

$SGS(D, q) = \{g_1, g_4, g_5, g_7\}$.

On peut facilement vérifier que g_2 (resp. g_3) $\notin$ *SGS(D, q)* car il est dominé par g_7 (resp. g_5) et $g_6 \notin$ *SGS(D, q)* car il est dominé par g_1. Ainsi, les graphes de *D* maximalement similaires à q sont g_1, g_4, g_5 et g_7.

	$\|SCM(g_i, q)\|$
(g_1, q)	4
(g_2, q)	4
(g_3, q)	4
(g_4, q)	3
(g_5, q)	5
(g_6, q)	5
(g_7, q)	6

TAB. 2 – *Informations sur* $|SCM(g_i, q)|$.

Distance / Paire de graphes	$Dist_{Ed}(g_i, q)$	$Dist_{SCM}(g_i, q)$	$Dist_{UG}(g_i, q)$
(g_1, q)	**4**	**0.33**	**0.50**
(g_2, q)	4	0.43	0.56
(g_3, q)	3	0.43	0.56
(g_4, q)	**2**	**0.50**	**0.67**
(g_5, q)	3	0.38	0.44
(g_6, q)	4	0.44	0.50
(g_7, q)	4	0.40	0.40

TAB. 3 – *Mesures de distance.*

6 Raffinement du skyline

Un des problèmes qui peut survenir lors du calcul de l'ensemble *SGS* (et d'un skyline en général) est sa taille qui est souvent très importante. D'un point de vue utilisateur, il est très souhaitable de disposer d'un critère pertinent pour sélectionner un sous-ensemble, de taille raisonnable, des graphes les plus intéressants parmi ceux de l'ensemble SGS. Une solution à ce problème est d'utiliser *le critère de diversité* (McSherry, 2002) pour sélectionner un sous-ensemble de graphes qui *est aussi divers que possible*, et ainsi fournir à l'utilisateur une image globale de l'ensemble *SGS*.

Soit *S* un sous-ensemble de *SGS*. La diversité de *S* signifie que les graphes qu'il contient doivent être dissimilaires. L'objectif est d'extraire de *SGS* un sous-ensemble S^* de taille *k* avec *une diversité maximale* (où *k* est un paramètre défini par l'utilisateur). Adaptée des travaux de McSherry (2002) et de ceux de Kukkonen et Lampinen (2007) et Hüllermeier et al. (2008), l'approche proposée définit la diversité de *S* ($\subseteq SGS$) de taille *k* par un vecteur $Div(S) = (v_1, v_2, v_3)$ tel que

$$v_i = min\{Dist_i(g, g') | g, g' \in S\},$$

où pour $i = 1, \ldots, 3$, $Dist_1 = Dist_{N-Ed}$ (i.e., version normalisée de $Dist_{Ed}$), $Dist_2 = Dist_{SCM}$ et $Dist_3 = Dist_{UG}$. La valeur v_i exprime la diversité sur la $i^{ème}$ dimension du sous-ensemble *S*.

Afin d'identifier le sous-ensemble S^*, nous considérons tous les sous-ensembles $S \subseteq SGS$ avec |S| = k comme candidats et appliquons les étapes suivantes :

Étape 1. Pour chaque dimension *i (i = 1, . . . , 3)*, ordonner d'une manière décroissante tous les sous-ensembles candidats *S* selon leur diversité v_i. Soit $rang_i(S)$ le rang de *S* par rapport à la $i^{ème}$ dimension. Rang de valeur *1* signifie la meilleure valeur de diversité et rang de valeur *M* signifie la plus mauvaise valeur de diversité (*M* est le nombre de sous-ensembles de taille *k* de l'ensemble *SGS*).

Étape 2. Évaluer un candidat *S* par :

$$val(S) = \sum_{i=1,\ldots,3} rang_i(S).$$

Le sous-ensemble minimisant ce critère (i.e., minimise la somme de ses positions dans tous les rangs) est considéré comme le sous-ensemble ayant la diversité maximale. Ainsi, S^* est caractérisé par

$$val(S^*) = min_S val(S),$$

où $S \subseteq SGS$ et |*S*| = *k*.

Exemple 5. Reprenons l'exemple donné en section 5 où $SGS(D, q) = \{g_1, g_4, g_5, g_7\}$. Supposons maintenant que l'utilisateur est intéressé par les *k (=2)* meilleurs graphes par rapport au critère de la diversité. On peut facilement vérifier que l'ensemble de tous les candidats contient *6* sous-ensembles de taille *k*, voir le tableau 4.

Diversité $v_{i,i=1,\ldots,3}$ / Candidat $S_{j,j=1,\ldots,6}$	v_1	v_2	v_3
$S_1 = \{g_1, g_4\}$	0.86	0.67	0.80
$S_2 = \{g_1, g_5\}$	0.83	0.50	0.60
$S_3 = \{g_1, g_7\}$	0.87	0.60	0.67
$S_4 = \{g_4, g_5\}$	0.80	0.62	0.73
$S_5 = \{g_4, g_7\}$	0.83	0.70	0.77
$S_6 = \{g_5, g_7\}$	0.75	0.50	0.61

TAB. 4 – *Candidats avec leur diversité.*

L'étape 1 et 2 conduisent aux résultats décrits dans le tableau 5. À partir du tableau 5-b, on peut facilement voir que $val(S_1)$ est la valeur minimale. Ainsi, $S^* = S_1 = \{g_1, g_4\}$ représente le sous-ensemble le plus divers possible du skyline $SGS(D, q)$.

7 Conclusion

Dans cet article, nous avons proposé une nouvelle approche permettant la recherche de graphes par similarité dans une BDG. Le concept clé de cette approche est la notion de *skyline de graphes*, que nous avons introduite. Ce type de skyline permet l'extraction de tous les graphes de la base de données interrogée qui ne sont pas dominés au sens de la relation de

	r_1	r_2	r_3
S_1	2	2	1
S_2	3	5	6
S_3	1	4	4
S_4	4	3	3
S_5	3	1	2
S_6	5	5	5

(a) Rangs ($r_i = rang_i$)

	$\sum_{i=1,\ldots,3} r_i$
S_1	5
S_2	14
S_3	9
S_4	10
S_5	6
S_6	15

(b) Val(S_i)

TAB. 5 – *Évaluation des Candidats.*

dominance par similarité suggérée, c'est-à-dire, les graphes qui sont maximalement similaires au graphe de la requête. Chaque graphe réponse est proposé à l'utilisateur avec un vecteur de scores montrant différentes similarités, correspondant à différentes caractéristiques, avec sa requête. Nous avons aussi montré comment sélectionner un sous-ensemble de taille acceptable du skyline et qui soit le plus divers possible. Une des perspectives de ce travail concerne l'évaluation de l'approche sur des données réelles afin de démontrer son efficacité et sa pertinence. Dans ce but, un prototype est en cours de réalisation.

Remerciement

Ce travail est co-financé par l'Agence Nationale pour la Recherche sous le projet AOC de référence ANR-08-CORD-009 et la région Bretagne. Nous tenons à leur exprimer nos vifs remerciements.

Références

Borzsonyi, S., D. Kossmann, et K. Stocker (2001). The skyline operator. In *Proc. of ICDE*, pp. 421–430.

Bunke, H. (1997). On a relation between graph edit distance and maximum common subgraph. *Pattern Recogn. Letters 18 (9)*, 689–697.

Bunke, H. et K. Shearer (1998). A graph distance metric based on the maximal common subgraph. *Pattern Recogn. Letters 19 (3-4)*, 255–259.

Chen, C., X. Yan, P. S. Yu, J. Han, D.-Q. Zhang, et X. Gu (2007). Towards graph containment search and indexing. In *Proc. of VLDB*, Vienna, Austria, pp. 926–937.

Hadjali, A., O. Pivert, et H. Prade (2010). Possibilistic contextual ckylines with incomplete preferences. In *Proc. of SoCPaR, Cergy Pontoise, Paris, France*.

He, H. et A. K. Singh (2006). Closure-tree: An index structure for graph queries. In *Proc. of ICDE*, pp. 38.

Hüllermeier, E., I. Vladimirskiy, B. P. Suárez, et E. Stauc (2008). Supporting case-based retrieval by similarity skylines: Basic concepts and extensions. In *Proc. of ECCBR*, pp. 240–254.

Khalefa, M. E., M. F. Mokbel, et J. J. Levandoski (2008). Skyline query processing for incomplete data. In *Proc. of ICDE*, pp. 556–565.

Kukkonen, S. et J. Lampinen (2007). Ranking-dominance and many-objective optimization. In *IEEE Congress on Evolutionary Computation*, pp. 3983–3990.

McSherry, D. (2002). Diversity-conscious retrieval. In *Proc. of ECCBR*, pp. 219–233. Springer-Verlag.

Pei, J., B. Jiang, X. Lin, et Y. Yuan (2007). Probabilistic skylines on uncertain data. In *Proc. of VLDB*, pp. 15–26.

Petrakis, E. G. M. et C. Faloutsos (1997). Similarity searching in medical image databases. *Proc. of TKDE 9 (3)*, 435 –447.

Shang, H., K. Zhu, X. Lin, Y. Zhang, et R. Ichise (2010). Similarity search on supergraph containment. In *Proc. of ICDE*, pp. 637–648.

Tian, Y. et J. M. Patel (2008). Tale : A tool for approximate large graph matching. In *Proc. of ICDE, Cancun, Mexico*, pp. 963–972.

Wallis, W. D., P. Shoubridge, M. Kraetz, et D. Ray (2001). Graph distances using graph union. *Pattern Recogn. Letters 22 (6-7)*, 701–704.

Yan, X., P. S. Yu, et J. Han (2004). Graph indexing: A frequent structurebased approach. In *Proc. of ACM SIGMOD*, pp. 335–346.

Yan, X., P. S. Yu, et J. Han (2005). Substructure similarity search in graph databases. In *Proc. of ACM SIGMOD*, pp. 766–777.

Yiu, M. L. et N. Mamoulis (2007). Efficient processing of top-k dominating queries on multi-dimensional data. In *Proc. of VLDB*, pp. 483–494.

Zhang, S., M. Hu, et J. Yang (2007). Treepi: A novel graph indexing method. In *Proc. of ICDE*, pp. 966–975.

Zhang, S., J. Z. Li, H. Gao, et Z. Zou (2009). A novel approach for efficient supergraph query processing on graph databases. In *Proc. of EDBT*, pp. 204–215.

Summary

One of the fundamental problems in graph databases is similarity search for graphs of interest. Existing approaches dealing with this problem rely on a single similarity measure between graph structures. In this paper, we suggest an approach allowing for searching similar graphs to a query graph where similarity between graphs is rather modelled by a vector of scalars than a unique scalar. To this end, we use the concept of *similarity skyline* of a query graph defined by the subset of graphs of the target database that are *the most similar* to the query in *a Pareto sense*. The idea is to achieve a d-dimensional comparison between graphs in terms of d local similarity measures and to retrieve those graphs that are *maximally similar* in the sense of a defined similarity dominance relation. A diversity-based method for refining the retrieval result is discussed as well.