Une Approche Skyline pour l'Interrogation de Bases de Données de Graphes

Katia Abbaci, Allel Hadjali, Ludovic Liétard, Daniel Rocacher

Une Approche Skyline pour l'Interrogation de Bases de Données de Graphes

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

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

Résumé. La recherche de graphes similaires à une requête à graphe est l'un des problèmes fondamentaux des bases de données de graphes. 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 permettant de rechercher les graphes similaires au graphe d'une requête où la similarité entre graphes n'est plus un scalaire unique mais un vecteur de scalaires. Pour cela, nous introduisons 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*. Une méthode pour raffiner le résultat de la recherche est aussi proposée en s'appuyant sur le critère de *diversité* entre les graphes.

1 Introduction

Les graphes sont devenus de plus en plus importants dans la modélisation des données structurées et complexes dans différentes domaines d'applications récentes : la bio-informatique (Hu et al., 2005), la reconnaissance de formes (Conte et al., 2004), les documents XML (Zhang et al., 2006), la chimie (Klinger et Austin, 2005), les réseaux sociaux (Cai et al., 2005), etc. Toutes ces applications indiquent l'importance et la large utilisation du paradigme des Bases de Données de Graphes (BDGs). D'une manière générale, on peut classer les requêtes adressées à une BDG en deux catégories (Zeng et al., 2009) : (1) *la recherche de graphes fondée sur une relation d'inclusion* et (2) *la recherche de graphes par similarité*. La première catégorie se compose en deux sous-problèmes suivants : (i) *la recherche de sous-graphes* : soit une BDG $D = \{g_1, g_2, \cdots, g_n\}$ 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) *la recherche de super-graphes* : soit une BDG $D = \{g_1, g_2, \cdots, g_n\}$ 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 graphe, 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; 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 structurellement similaires au graphe de la requête, est apparue comme une nouvelle tendance pour les raisons suivantes. Premièrement, de nombreuses et réelles BDGs sont de nature bruitée et incomplète, d'où la nécessité d'un *appariement approximatif* de graphes. Deuxièmement, plusieurs applications modernes préfèrent les résultats d'un appariement approximatif plutôt que ceux d'un appariement exact car ils véhiculent davantage d'informations, comme ce qui pourrait être manquant ou superflu dans un graphe de requête ou dans une BDG. Ainsi, plusieurs approches ont été proposées pour répondre aux requêtes de recherche de graphes par similarité (ou simplement, requêtes par similarité). Voir (Yan et al., 2005; He et Singh, 2006; Tian et Patel, 2008; Shang et al., 2010).

Le point commun entre toutes ces approches est l'utilisation d'une *seule* mesure pour évaluer la similarité entre graphes. Toutefois, un graphe est une structure complexe 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 évaluer d'une manière significative et efficace la similarité entre graphes. Chaque indice est dédié à mesurer une distance (ou similarité) locale afférent à un aspect donné dans la structure du graphe. Ainsi, la similarité entre graphes est caractérisée par un vecteur de mesures de distances locales au lieu d'une seule mesure. De cette façon, on peut préserver l'information concernant la similarité sur différentes caractéristiques, lorsque l'on compare deux graphes.

Nous proposons une nouvelle approche de traitement de requêtes par similarité en utilisant la notion de *skyline par similarité*. 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 *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 distances locales et d'identifier les graphes qui sont *maximalement similaires* au sens d'une relation de *dominance par similarité*. Les principales contributions de l'article sont :

1. Nous introduisons la notion de *similarité composée entre graphes* (SCG) et définissons ensuite la *relation de dominance par similarité entre graphes*.
2. En se basant sur cette relation de dominance, nous proposons une définition formelle du *skyline par similarité entre graphes*, i.e., les graphes de la base interrogée maximalement similaires au graphe de la requête au sens de Pareto.
3. Pour réduire la taille du skyline, qui est souvent très importante, nous proposons une méthode permettant d'extraire un sous-ensemble de graphes qui est *aussi divers que possible*, mais dont la taille est raisonnable.

L'article est structuré comme suit. La section 2 introduit quelques notions préliminaires. La section 3 discute quelques travaux apparentés. La section 4 décrit des mesures de calcul de similarité entre graphes déjà existantes et leur interprétations. Dans la section 5, nous introduisons la notion de skyline par similarité dédiée aux requêtes à graphe. La section 6 présente un exemple détaillé. Dans la section 7, nous proposons une méthode pour raffiner les résultats retournés dans le skyline. La section 8 conclut l'article.

2 Notions préliminaires

2.1 Rappel sur les requêtes skyline

Les requêtes skyline (Borzsonyi et al., 2001) représentent un paradigme très populaire et puissant pour extraire des objets d'un ensemble de données multidimensionnel. 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 domine (est préféré à) q* et on note $p \succ q$.

Définition 2. Le skyline de *r* est le sous-ensemble des points non-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 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

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 et H_3, car H_1 est dominé par H_2.

2.2 Quelques définitions de base

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*.

Par souci de clarité, les graphes considérés sont étiquetés et non-orientés (différents nœuds peuvent avoir la même étiquette). La taille d'un graphe *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 graphes). 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 → 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-graphes). Soit deux graphes *g = (V, E, L, l)* et *g' = (V', E', L', l')*, *g* est *isomorphe de sous-graphes à g'* s'il existe une injection *f : V → 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-graphes 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 Travaux apparentés

L'étude présentée dans cet article peut être apparentée avec les travaux effectués dans les domaines des requêtes skyline et des requêtes de recherche de graphes par similarité.

Requêtes skylines. Elles ont reçu l'attention de nombreux chercheurs. Plusieurs études ont été menées pour développer des algorithmes efficaces et introduire des variantes pour les requêtes skyline (Pei et al., 2007; Yiu et Mamoulis, 2007; Khalefa et al., 2008; Hadjali et al., 2010). Pour autant que nous le sachions, il n'existe pas de travaux portant sur les requêtes skyline dans un contexte d'interrogation de BDGs, excepté, le travail de Zou et al. (2010) qui étudie les requêtes skyline dynamique dans le cadre d'un graphe de grande taille. Dans notre cas, c'est un autre type de skyline (i.e. skyline par similarité) sur un ensemble de graphes qui est étudié.

Requêtes par similarité. Plusieurs approches ont été développées pour traiter les requêtes par similarité. *Grafil* (Yan et al., 2005) applique une recherche par similarité de sous-structures dans une BDG à large échelle. Il retourne tous les graphes de la base de données qui contiennent approximativement le graphe de la requête. *C-Tree* (He et Singh, 2006) est un autre outil pour la recherche par similarité de sous-graphes. Il est basé sur la distance d'édition entre la requête et les graphes candidats. *Tale* (Tian et Patel, 2008) propose une technique d'appariement innovante qui distingue les nœuds par leur importance dans la structure du graphe. Cette technique met d'abord en correspondance les nœuds importants d'une requête à graphe, ensuite, elle étend progressivement ces correspondances. Récemment, Shang et al. (2010) ont proposé une technique répondant aux requêtes super-graphes où le problème de recherche par similarité de super-graphes est converti en un problème de détection de σ-sous-graphes manquants, où σ est un seuil de tolérance d'erreurs. Les deux approches *C-Tree* et *Tale* utilisent la distance d'édition pour mesurer la similarité entre graphes, tandis que les travaux réalisés par Yan et al. (2005) et Shang et al. (2010) utilisent la notion de *sous-graphe commun maximal*.

Comme on peut le constater, toutes les approches proposées et dédiées aux requêtes par similarité utilisent un seul indice pour mesurer la similarité entre deux graphes. En procédant ainsi, la similarité entre deux graphes n'est pas entièrement capturée car des similitudes relatives à certaines caractéristiques du graphe pourraient être ignorées. Ceci est principalement dû au fait que chaque indice de similarité entre graphes peut être vu comme une mesure locale qui exprime seulement une ressemblance au regard d'un seul aspect dans la structure des graphes (voir la section 4). Par comparaison avec les travaux ci-dessus, notre approche, d'une part, repose sur une mesure de similarité composée entre graphes, et d'autre part, retourne un ensemble de graphes dominants par similarité au sens de Pareto pour répondre à une requête.

4 Similarité de graphes : une vue sémantiques

Plusieurs modèles ont été proposés (Bunke, 1997; Bunke et Shearer, 1998; Wallis et al., 2001) pour mesurer la similarité (ou la distance) entre deux graphes. Ci-après, nous présentons les mesures les plus utilisées pour déterminer les similarités entre graphes[1].

4.1 La distance d'édition de graphes

La distance d'édition de graphes (Bunke, 1997; He et Singh, 2006) se base sur les opérations d'édition de graphes nécessaires pour transformer un graphe en un autre. Généralement, l'ensemble d'opérations d'édition considéré inclut : l'insertion, la suppression et le ré-étiquetage de nœuds/d'arêtes. Chaque opération d'édition est associée à un coût (un nombre réel non négatif) en fonction de l'intensité de distorsion induite par la transformation. Soit *e_op* une opération d'édition et *c(e_op)* son coût. Le coût d'une séquence d'opérations d'édition, $s = e\_op_1, \ldots, e\_op_n$ est donné par

$$c(s) = \sum_{i=1}^{n} c(e\_op_i).$$

En pratique, il est difficile de déterminer le coût de chaque opération d'édition élémentaire. Par souci de simplicité, on considère, ici, coût est égale à une mesure de distance *uniforme* : la distance entre deux nœuds/arêtes est *1* si leurs étiquettes sont différentes ; *0* sinon.

Définition 8 (Distance d'édition de graphes). La distance d'édition entre deux graphes g_1 et g_2 est égale au *coût minimal*, résultante de toutes les séquences d'opérations d'édition possibles, qui transforment g_1 en g_2, i.e.,

$$Dist_{Ed}(g_1, g_2) = min_{s \in E\_op} c(s) \quad (1)$$

où *E_op* dénote l'ensemble de toutes les séquences d'opérations d'édition possibles qui transforment g_1 en g_2. Plus $Dist_{Ed}(g_1, g_2)$ est faible, plus les deux graphes sont similaires. Cette mesure de distance s'applique à tout type de graphes sans aucune restriction.

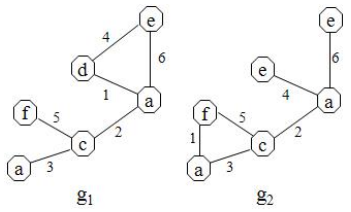

FIG. 1 – *Exemple de graphes étiquetés.*

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

[1] Par souci d'espace, l'analyse de la complexité du calcul de chaque mesure n'est pas abordée dans cet article.

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, $Dist_{Ed}(g_1, g_2) = 4$.

Dans le contexte d'interrogation de BDGs, cette mesure de distance nous renseigne sur les caractéristiques non partagées par un graphe cible de la BDG et le graphe de la requête.

4.2 La distance basée sur le SCM

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

Définition 9 (Similarité basée sur le SCM). Soit deux graphes g_1 et g_2, la similarité entre graphes 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 $|SCM(g_1, g_2)|$ dénote le nombre d'arêtes dans $SCM(g_1, g_2)$.

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 entre graphes, $Dist_{SCM}$, dérivée de Sim_{SCM} peut s'écrire :

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

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.

Example 3. Reprenons l'exemple 2. La distance basée sur le *SCM* entre g_1 et g_2 est calculée comme suit. Premièrement, le *SCM* (g_1, g_2) est identifié, voir la figure 2. Ensuite, par application de (2), nous obtenons

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

où $|SCM(g_1, g_2)| = 4$ et $max(|g_1|, |g_2|) = 6$.

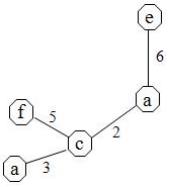

FIG. 2 – *SCM de g_1 et g_2.*

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

4.3 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.

Définition 10 (Similarité basée sur l'UG). Soit deux graphes g_1 et g_2, la similarité entre 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[2].

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 (Wallis et al., 2001) 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 :

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

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

$Dist_{UG}(g_1, g_2) = 1 - \frac{|SCM(g_1,g_2)|}{|g_1|+|g_2|-|SCM(g_1,g_2)|} = 0.50,$

où $|SCM(g_1, g_2)| = 4$ (voir l'example 3) et $|g_1| = |g_2| = 6$.

Dans le cadre d'interrogation de BDGs, cette similarité donne également des informations sur les aspects communs entre un graphe de la base interrogée et le graphe de la requête.

5 Skyline par similarité entre graphes

Dans ce qui suit, nous supposons que la similarité entre graphes est une notion composée, i.e., un vecteur de mesures de distance.

Définition 11 (Similarité composée entre 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

$$SCG(g, g') = (Dist_1(g, g'), Dist_2(g, g'), \ldots, Dist_d(g, g')),$$

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

Soit $D = \{g_1, g_2, \ldots, g_n\}$ une BDG et q une requête par similarité. Pour répondre à q, 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 à q au sens de la relation de dominance par similarité définie ci-dessous.

Définition 12 (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. $\exists k \in \{1, \ldots, d\}, Dist_k(g, q) < Dist_k(g', q)$.

[2]Cette mesure de similarité ressemble à l'indice de Jaccard utilisé pour mesurer la similarité entre deux ensembles *A* et *B*, i.e., $J(A, B) = |A \cap B| \,/\, |A \cup B|$.

Plus simplement, 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 la 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)* :

$$SGS(D, q) = \{g \in D | \nexists g' \in D, g' \succ_q g\} \quad (4)$$

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

Pour illustrer notre 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 de mesures de distance locales décrites dans la section 4, i.e.,

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

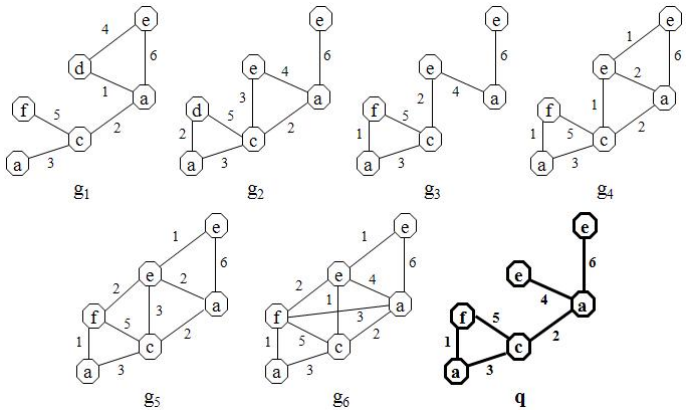

FIG. 3 – *La base de données D et le graphe requête q.*

6 Un exemple illustratif

Soit $D = \{g_1, g_2, g_3, g_4, g_5, g_6\}$ une BDG et q un requête par similarité (voir 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)*. Les valeurs de $|SCM(g_i, q)|$, pour i =1, . . .,6, sont données dans le tableau 2 et les vecteurs de similarité entre graphes $SCG(g_i, q)$, pour i=1, . . .,6, sont résumés dans le tableau 3. Par application de (4), l'ensemble des graphes optimaux au sens de Pareto, i.e. le skyline de graphes par similarité, est donné par $SGS(D, q) = \{g_1, g_3, g_4, g_6\}$.

Il est aisé de voir que $g_2 \notin SGS(D, q)$ car il est dominé par g_4 et $g_5 \notin SGS(D, q)$ car il est dominé par g_1. Les graphes de D maximalement similaires à q sont g_1, g_3, g_4 et g_6. En effet,

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

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

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

TAB. 3 – *Mesures de distance.*

- Le graphe g_1 est le plus intéressant au regard de la mesure $Dist_{SCM}$. Cela est dû aux raisons suivantes : i) g_1 satisfait un maximum de caractéristiques requises par q que d'autres graphes de même taille ; ii) g_1 et q sont de même taille. Mais, g_1 est le moins intéressant au regard des caractéristiques manquantes et superflues (i.e., $Dist_{Ed}$).
- Le graphe g_3 est le meilleur au regard de la mesure $Dist_{Ed}$. Cela signifie, d'une part, qu'il est le plus intéressant au regard du nombre de désaccords avec q. D'autre part, g_3 est beaucoup moins satisfaisant au regard des concordances avec q au sens de *SCM*.
- Le graphe g_6 est le plus intéressant au regard de la mesure $Dist_{UG}$. Cela est dû au fait que $g_6 \supset q$. Mais, c'est le moins intéressant au regard du critère basé sur les caractéristiques superflues (i.e., $Dist_{Ed}$).
- Le graphe g_4 peut être considéré comme un bon compromis entre les trois mesures $Dist_{Ed}$, $Dist_{SCM}$ et $Dist_{UG}$.

Examinons maintenant les résultats obtenus en utilisant seulement une seule mesure de similarité entre graphes. Si on s'intéresse aux *k* (= 3) meilleures réponses, g_2 est retourné comme graphe candidat en utilisant l'approche basée sur la distance d'édition. Tandis qu'avec l'approche basée sur le skyline, g_2 n'est pas retourné à l'utilisateur car g_4 est meilleur.

7 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 permettant de sélectionner un sous-ensemble, de taille raisonnable, des graphes les plus intéressants parmi ceux du skyline 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 des éléments de *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 à partir de *SGS* un sous-ensemble $\mathbb{S}$ de taille *k* (*k* est un paramètre défini par l'utilisateur) avec *une diversité maximale*. En s'inspirant des travaux de Kukkonen et Lampinen (2007), 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, ..., 3, $Dist_1 = Dist_{Ed}$, $Dist_2 = Dist_{SCM}$ et $Dist_3 = Dist_{UG}$. La valeur v_i exprime la diversité sur la i^{me} dimension du sous-ensemble *S*.

Afin d'identifier le sous-ensemble $\mathbb{S}$, 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* au regard de 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*).

Etape 2. Évaluer un candidat *S* par : $val(S) = \sum_{i=1,...,3} rang_i(S)$.

Le sous-ensemble minimisant la somme de ses positions dans tous les rangs est considéré comme le sous-ensemble ayant la diversité maximale. Ainsi, $\mathbb{S}$ est caractérisé par

$val(\mathbb{S}) = min_S val(S)$, où $S \subseteq SGS$ et $|S| = k$.

Exemple 5. Reprenons l'exemple donné dans la section 6 où le skyline $SGS(D, q) = \{g_1, g_3, g_4, g_6\}$. Supposons maintenant que l'utilisateur est intéressé par les *k (=2)* meilleurs graphes au regard du 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.

	v_1	v_2	v_3
$S_1 = \{g_1, g_3\}$	0.86	0.67	0.80
$S_2 = \{g_1, g_4\}$	0.83	0.50	0.60
$S_3 = \{g_1, g_6\}$	0.87	0.60	0.67
$S_4 = \{g_3, g_4\}$	0.80	0.62	0.73
$S_5 = \{g_3, g_6\}$	0.83	0.70	0.77
$S_6 = \{g_4, g_6\}$	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 duquel on peut voir que $val(S_1)$ est la valeur minimale. Ainsi, $\mathbb{S} = S_1 = \{g_1, g_3\}$.

8 Conclusion

Dans cet article, nous avons proposé une approche permettant la recherche de graphes par similarité. Le concept clé de cette approche est la notion de *skyline de graphes par similarité*.

	r_1	r_2	r_3	$Val(S_i) = \sum_{i=1,\cdots,3} r_i$
S_1	2	2	1	5
S_2	3	5	6	14
S_3	1	4	4	9
S_4	4	3	3	10
S_5	3	1	2	6
S_6	5	5	5	15

TAB. 5 – *Évaluation des Candidats ($r_i = rang_i$).*

Ce type de skyline permet l'extraction de tous les graphes de la base de données interrogée qui ne sont dominés par aucun autre graphe de la base au sens de la relation de dominance par similarité définie. Chaque graphe réponse est retourné à l'utilisateur avec un vecteur de scores montrant les différentes similarités correspondant aux différentes caractéristiques avec le graphe de la requête. Nous avons aussi montré comment sélectionner un sous-ensemble de diversité maximale à partir d'un skyline de graphes par similarité. Nous travaillons actuellement sur l'implémentation de l'approche pour démontrer son efficacité et sa pertinence.

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.

Cai, D., Z. Shao, X. He, X. Yan, et J. Han (2005). Community mining from multirelational networks. In *Proc. of PPKDD*, pp. 445–452.

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.

Conte, D., P. Foggia, C. Sansone, et M. Vento (2004). Thirty years of graph matching in pattern recognition. *Inter. J. of Pattern Recogn. and Art. Intell. 18 (3)*, 265–298.

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

He, H. et A. K. Singh (2006). Closure-tree: An index structure for graph queries. In *Proc. of ICDE*, pp. 38–54.

Hu, H., Y. Hang, J. Han, et X. Zhou (2005). Mining coherent dense subgraphs across massive biological network for fonctional discovery. *Bioinformatics 1(1)*, 1–9.

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

Klinger, S. et J. Austin (2005). Chemical similarity searching using a neural graph matcher. In *Proc. of ESANN*, pp. 479–484.

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.

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

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 (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.

Zeng, Z., A. Tung, J. Wang, J. Feng, et L. Zhou (2009). Comparing stars: On approximating graph edit distance. In *Proc. of VLDB*, pp. 25–36.

Zhang, N., T. Özsu, I. Ilyas, et A. Aboulnaga (2006). Fix: Feature-based indexing technique for xml documents. In *Proc. of VLDB*, pp. 259–270.

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.

Zou, L., L. Chen, M. T. Ozsu, et D. Zhao (2010). Dynamic skyline queries in large graphs. In *Proc. of DASFAA*, pp. 62–78.

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 introduce 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*. A diversity-based method for refining the retrieval result is proposed as well.