M-Qube

une méthode d'exploration itérative par extractions successives de vues partielles

Sommaire

Dans le chapitre précédent, nous avons explicité les spécificités des sciences humaines et sociales. A partir de cela, nous dégageons les éléments auxquels notre méthode doit apporter une solution :

- Une attention centrée essentiellement sur les individus et leurs interconnexions en réduisant l'échelle du graphe à un objet lisible et analysable par l'utilisateur sans passer par une agrégation ou une simplification des données.
- Une méthode de travail nécessitant de créer en parallèle une arborescence de pistes dans lesquelles il est nécessaire de pouvoir naviguer.

- Une gestion du caractère multi-couche des réseaux en sciences humaines et sociales afin de pouvoir différencier les traitements en fonction des types de données considérées.

Nous introduisons ainsi `M-QuBE`[3] (**M**ultilayer network : **Qu**erying **B**ig networks by **E**volutive **E**xtraction and **E**xploration), une méthode permettant la construction d'une succession arborescente de sous-réseaux pour convenir à la méthodologie des experts des données, en offrant la possibilité, à n'importe quel moment du processus, d'affiner le chemin suivi jusqu'à présent pour parfaire son analyse, de revenir à un état antérieur pour essayer d'autres chemins ou simplement de décider d'explorer de nouvelles pistes découlant d'essais antérieurs.

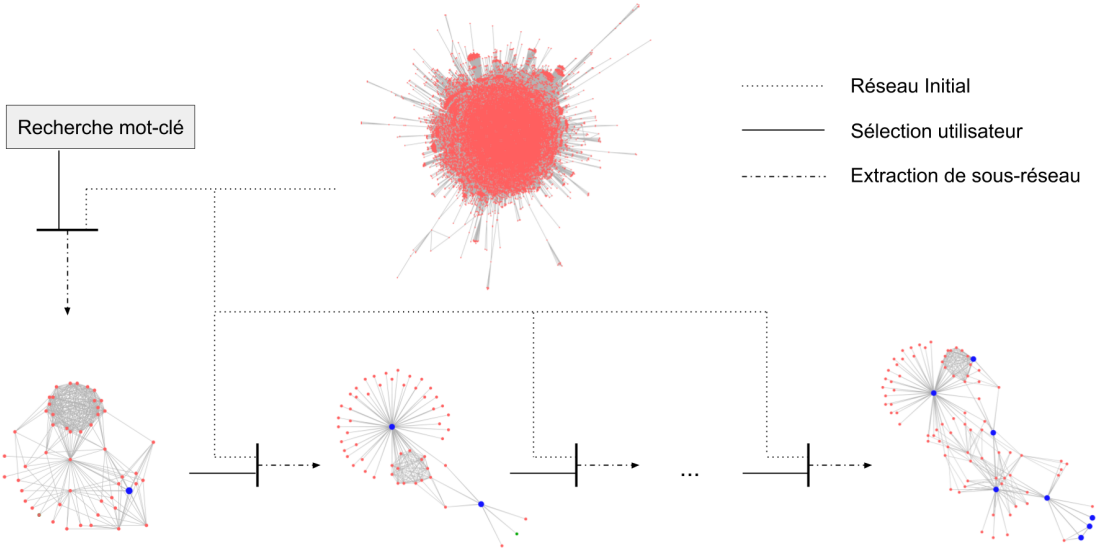


FIGURE *4.1* – ***Fonctionnement itératif du processus*** *: Création d'une série de sous-graphes. Les sommets bleus sont sélectionnés par l'utilisateur. Chaque sélection utilisateur permet la création d'un nouveau sous-graphe d'intérêt supérieur. Ce dernier est issu du graphe initial et prend en compte les informations sélectionnées dans les sous-graphes précédents.*

Le fonctionnement général de ce processus, comme illustré dans la Fig. 4.1, consiste à utiliser l'interaction de l'utilisateur (sélection de sommet ou recherche par mot-clé) pour créer, à partir du réseau initial, une succession potentiellement arborescente de sous-réseaux de plus en plus pertinents pour l'utilisateur.

Présentons une illustration concrète de cette méthode (Fig. 4.2) utilisant les données du CVCE (et la configuration décrite dans la section 6.1.3). Dans ce

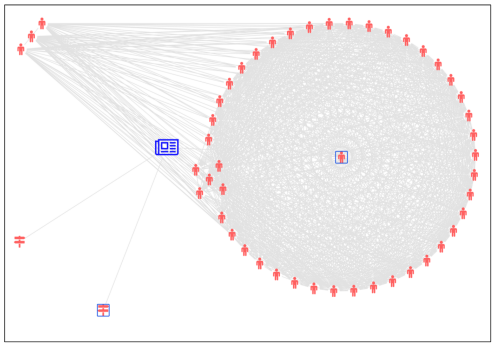
scénario, nous souhaitons évaluer l'influence du Royaume-Uni dans l'histoire de la construction européenne. Notre point de départ est la photo d'un meeting ayant eu lieu à Londres entre Margaret Thatcher et Helmut Schmidt témoignant d'une interaction entre le Royaume-Uni et un représentant de l'Union européenne. Un premier sous-réseau est généré à partir de la sélection de ce document (Fig. 4.2a). Dans ce sous-réseau, nous sélectionnons M.Thatcher et Londres pour valoriser les éléments du réseau en lien avec l'une des plus célèbres politiciennes anglaises ou un lieu phare du Royaume-Uni.

Nous obtenons un nouveau sous-réseau avec un nouveau contexte cette fois centré sur Margaret Thatcher, Londres et le document original. Nous constatons l'apparition d'un nouveau document sur l'entrée d'un nouveau pays dans l'Union européenne et de nombreuses personnalités politiques étant liées à ce document. Parmi eux figurent Jacques Delors et Pierre Werner qui sont liés à la fois à Margaret Thatcher ainsi qu'à plusieurs autres personnes en lien avec elle. Nous les sélectionnons afin d'orienter le contexte sémantique vers davantage d'informations relatives aux acteurs majeurs de l'Europe afin de pouvoir les corréler avec les informations relatives à M.Thatcher (Fig. 4.2b).

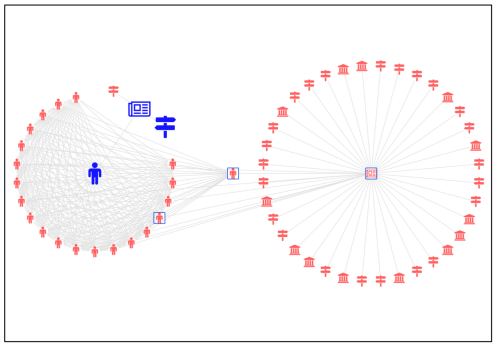
Le nouveau sous-réseau généré fournit une large gamme de nouvelles informations (Fig. 4.2c). Des documents relatifs aux trois acteurs sélectionnés apparaissent et concernent des sujets directement liés à notre objectif tels que le référendum anglais de 1975 ou des commentaires sur la vision de l'Europe de M.Thatcher. On peut également noter que ces documents estimés pertinents sont liés à des institutions européennes telles que la Commission européenne ou la Communauté économique européenne (CEE) qui figurent également dans ce nouveau sous-réseau. La sélection de ces entités va permettre d'orienter les documents et les personnalités et donc de faire évoluer à nouveau le contexte sémantique.

Nous décidons plutôt d'explorer une nouvelle voie car notre curiosité nous pousse à faire évoluer notre objectif initial. Pour ce faire, nous ne retenons de la sélection actuelle que M.Thatcher et nous sélectionnons un sommet représentant la République française, le Président français ainsi qu'un article francophone très critique envers l'Europe. Le nouveau sous-réseau généré (Fig. 4.2d) est entièrement construit à partir de la nouvelle sélection et offre ainsi un nouvel horizon de recherche en accord avec cette nouvelle voie.

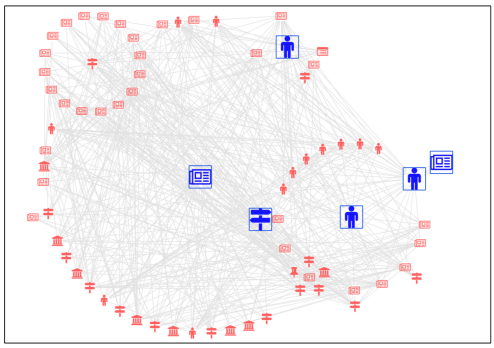
Avec `M-QuBE`^{3}, les experts sont donc en mesure d'explorer et de guider leurs explorations à travers un large réseau simplement en analysant les successions de sous-réseaux de tailles réduites. Outre le processus de génération de sous-réseaux et les interactions utilisateur qu'il utilise, il est aussi nécessaire de proposer aux



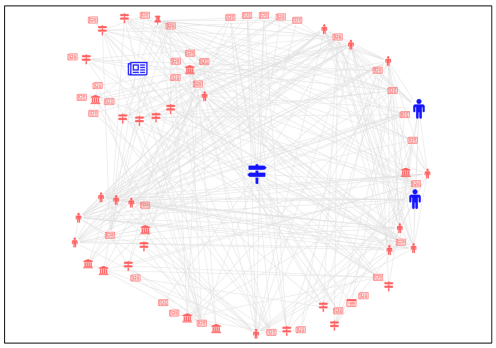
(a) Commençons notre exploration à partir d'un document sur un meeting européen à Londres (icône journal bleue). Nous sélectionnons Margaret Thatcher (icône personne encadrée en bleu) et Londres (icône pancarte encadrée en bleu) dans ce premier sous-réseau.



(b) Un second sous-réseau est généré à partir de la sélection réalisée en a) (les trois sommets agrandis bleus). Dans ce nouveau sous-réseau, nous sélectionnons un document sur l'entrée dans l'Europe d'un nouveau pays ainsi que deux acteurs majeurs de la scène européenne (Jacques Delors et Pierre Werner), tous liés à M.Thatcher et encadrés en bleu.



(c) Une nouvelle piste est explorée en ne conservant de la sélection actuelle uniquement Margaret Thatcher et en sélectionnant à la place les sommets représentant le rôle de président français, la république de France et un document traitant de la "Décadence Européenne".



(d) Ce nouveau sous-réseau présente un horizon de recherche entièrement nouveau calculé à partir de cette nouvelle sélection. Il est possible d'accéder à de nouveaux documents, nouvelles localisations, nouvelles institutions et nouvelles personnes liés directement ou indirectement à notre nouvelle sélection.

FIGURE *4.2 – Le rôle du Royaume-Uni dans le développement européen*

utilisateurs un moyen d'interagir et de naviguer dans la succession arborescente de sous-réseaux générés correspondant aux différentes pistes ayant été suivies (où "arbre de trace").

La méthode M-QuBE^3 se divise donc en deux parties : d'une part le processus de génération des sous-réseaux (Section 4.1), d'autre part la gestion de cette succession arborescente de sous-réseaux (Section 4.2).

4.1 Génération itérative de sous-réseaux

Le processus de génération de sous-réseaux de M-QuBE^3 est lui-même divisé en trois phases distinctes (Fig. 4.3). Le processus commence par une recherche par mot clé (partie A) afin de sélectionner des sommets. Ces sommets composent l'*ensemble focus* initial : une liste de sommets de référence permettant de définir une liste de sommets candidats qui seront potentiellement affichés dans le prochain sous-graphe à extraire. Un score est calculé pour ces sommets candidats en considérant les informations sémantiques des données ainsi que la structure du réseau afin d'obtenir une estimation d'intérêt pour l'utilisateur (partie B). A partir de ces scores, un classement est effectué pour déterminer une liste des sommets les plus intéressants (liste des élus) qui sont ensuite utilisés pour extraire le sous-réseau qui sera présenté à l'utilisateur (partie C).

Les experts commençant avec un détail et progressant pas à pas, le processus est structuré de manière similaire. Les étapes de sélection et d'extraction peuvent donc être répétées itérativement afin d'explorer les données plus en détail et avec une précision croissante.

L'utilisateur interagit avec le sous-graphe extrait en sélectionnant de nouveaux sommets qui lui semblent pertinents. Ces sommets vont venir enrichir l'ensemble focus et ainsi améliorer le prochain sous-réseau. Cet ensemble est conservé à travers les itérations et utilisé pour le calcul du prochain sous-réseau. Nous détaillons dans la suite les différentes étapes suivies pour chaque itération du processus.

Dans les parties suivantes, les figures représentant en détail les parties A, B et C (Fig. 4.4, 4.5 et 4.6) sont toutes trois extraites de la figure globale de M-QuBE^3 (Fig. 4.7). Les positionnements des éléments au sein des figures sont donc choisis afin de pouvoir être rassemblés dans la figure globale.

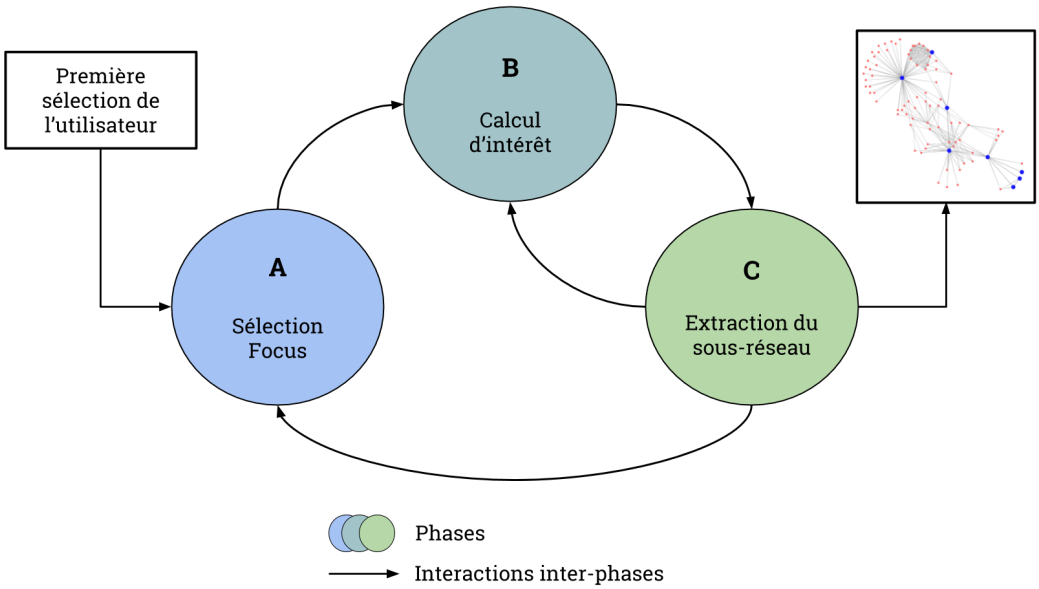


FIGURE *4.3* – ***Interactions entre les différentes phases du processus de génération de sous-réseaux de*** `M-QuBE`^{3} *: Pour commencer le processus, une première sélection est effectuée. A partir de cette sélection (A), plusieurs calculs d'intérêt vont être effectués dans le réseau (B) afin d'établir un classement entre les sommets et déterminer le plus intéressant pour l'utilisateur (C). Le sommet le plus intéressant est conservé dans un sous-réseau et, si celui-ci n'est pas de taille suffisante, un calcul est ré-itéré (B) en considérant de nouveaux éléments. Lorsque le sous-réseau est de taille satisfaisante pour l'utilisateur, celui-ci est montré et l'utilisateur peut alors ré-itérer le processus complet en établissant une nouvelle sélection (A).*

4.1.1 Sélection de l'ensemble focus

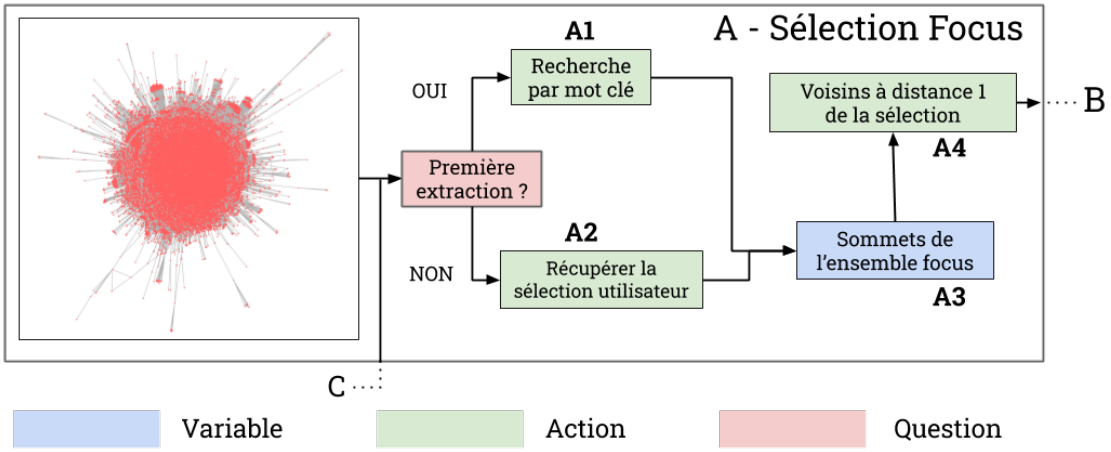


FIGURE *4.4* – ***Phase de sélection de l'ensemble focus*** *: La phase de sélection permet de définir l'ensemble focus, un ensemble de sommets de référence à partir duquel les sommets à évaluer sont déterminés. S'il s'agit du premier passage dans la phase A, la sélection s'effectue par recherche de mot-clé. Le cas échéant, la sélection sera définie par l'utilisateur en interagissant avec la vue d'un sous-réseau (par une zone de sélection ou un clic par exemple). Le voisinage des sommets focus sont ensuite transmis à la phase B.*

Comme énoncé précédemment, l'ensemble focus est l'ensemble des sommets sélectionnés par l'utilisateur. Dans un premier temps, il doit être initialisé ou être mis à jour. S'il s'agit de la première itération, l'ensemble focus (Fig. 4.4, A3) est défini par une recherche par mot clé d'un ou plusieurs sommets (A1). Ensuite, entre chaque nouvelle itération, l'ensemble focus est modifié par l'ajout ou la suppression de sommets par l'utilisateur (A2). Les voisins de chaque sommet de l'ensemble focus (A4) composent ensuite un ensemble appelé la liste de candidats (voir section suivante : Fig. 4.5, B2). Cette liste contient les sommets potentiellement montrés dans le prochain sous-réseau extrait, en fonction de l'estimation d'intérêt calculée dans le partie B. Cette liste de candidats est amenée à évoluer au cours de processus.

La prochaine phase est le calcul des différentes métriques utilisées pour déterminer le score d'intérêt final.

4.1.2 Calcul d'intérêt

La phase de calcul d'intérêt détermine un score (Fig. 4.5, B7) représentant l'intérêt de l'utilisateur pour un sommet donné n. Plus le score est élevé, plus la probabilité que le sommet n soit sélectionné et montré à l'utilisateur dans le prochain sous-réseau est grande (Fig. 4.6, C6).

Le score final est une métrique définie à partir de plusieurs scores. Premièrement, un score directement lié à la volonté utilisateur, l'*eScore* (B3), ainsi qu'un score défini en fonction de la position des sommets dans le réseau, *pScore* (B4), sont calculés. Ces deux scores sont ensuite combinés en un score pondéré, le *wScore* (B5). Le score final est finalement obtenu en prenant en compte les scores des voisins (B3' à B5') à travers un calcul de diffusion de l'intérêt (B6).

eScore (B3). L'eScore (**e**xploratory **Score**) est une métrique d'estimation de l'intérêt calculée pour tous les sommets dans la liste de candidats (B1). L'objectif est de représenter la volonté générale de l'utilisateur notamment les contraintes et les objectifs à prendre en compte lors de la recherche. Lors d'une première utilisation de `M-QuBE`^{3} sur un nouveau jeu de données ou en cas d'objectifs imprécis ou encore non-définis de la part de l'utilisateur, il est tout à fait possible d'utiliser des métriques communément utilisées en sciences humaines et sociales (comme les exemples énoncés dans les travaux de Mainas [54] appliqués aux réseaux criminels). Notre solution spécialisée d'eScore pour les réseaux multi-couches est présentée dans la section 5.2 du chapitre 5 pour adapter précisément `M-QuBE`^{3}aux réseaux issus des sciences humaines et sociales.

pScore (B4). L'utilisateur interagit avec le processus en sélectionnant des sommets à chaque itération (Partie A, Fig. 4.4). Nous supposons que les sommets proches d'un sommet sélectionné dans le graphe ont plus de chances d'être considérés comme intéressants par l'utilisateur. A cette fin, la sélection de l'utilisateur constitue ce que nous appelons une zone focale et un sommet inclus ou proche de cette zone est pondéré positivement (voir le calcul du score pondéré ci-dessous). Pour ce faire, une fonction basée sur le centroïde est utilisée pour calculer la distance moyenne entre le sommet évalué x et les sommet de l'ensemble focus, déterminant ainsi sa position dans la zone focale.

Cette fonction est définie ainsi :

$$C(x, Y) = \frac{\sum_{y \in Y} d(x, y)}{|Y|}$$

avec Y l'ensemble focus et d une fonction de distance. La fonction la plus pertinente pour d indépendamment du contexte est souvent un calcul du plus court chemin entre deux sommets. La distance euclidienne peut également être utilisée, mais les coordonnées des sommets issues de l'algorithme de dessin du graphe doivent avoir un sens, ce qui nécessite d'abord un travail sur le modèle et la représentation du réseau.

$pScore(x, Y)$ est la distance normalisée entre le sommet x et les sommets de l'ensemble Y :

$$pScore(x, Y) = 1 - \frac{C(x, Y) - c_{min}}{c_{max} - c_{min}}$$

où c_{min} et c_{max} sont respectivement la valeur minimum et maximum de C dans le réseau. Une normalisation est ici nécessaire pour pouvoir effectuer l'étape suivante en mettant à la même échelle les informations sémantiques (eScore) et les informations topologiques (pScore).

wScore (B5). Les métriques $eScore$ et $pScore$ sont combinées pour obtenir le score pondéré $wScore$ défini tel que :

$$wScore(x, Y) = (1 - w) \times eScore(x|Y) + w \times pScore(x, Y)$$

où w est une constante définie sur $[0; 1]$ en fonction de l'importance que veut donner l'utilisateur à la zone focale. $wScore$ est donc l'estimation d'intérêt d'un sommet pour l'utilisateur en tenant compte à la fois de la sémantique (eScore) et des informations structurelles du réseau (pScore).

Diffusion (B6). Le calcul de l'intérêt des sommets se termine par la phase de diffusion. Un problème possible de ce processus est le même que celui rencontré dans les travaux de van Ham et Perer [76]. La liste des sommets candidats s'étend itérativement à la manière d'un algorithme glouton : lorsqu'un sommet est sélectionné pour être montré dans le sous-réseau, ses sommets voisins sont ajoutés à la liste de candidats. Cependant, si un sommet très intéressant (un sommet avec un score élevé) est entouré de sommets avec un score faible, l'algorithme itératif

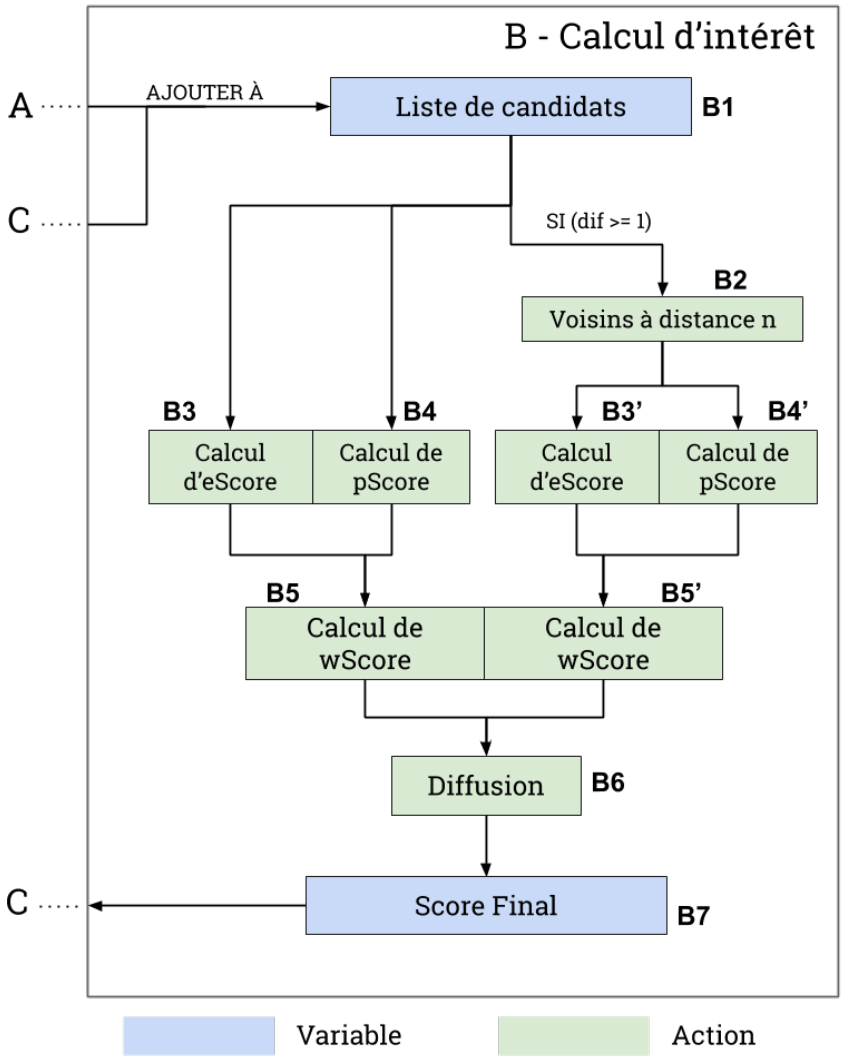


FIGURE 4.5 – ***Phase de calcul de l'intérêt*** *: Cette phase commence en initialisant une liste de candidat à partir des sommets reçus (B1). Si ceux-ci n'ont jamais été évalués, des scores basés sur des fonctions d'intérêt (B3) et sur la position des sommets dans le réseau (B4) sont calculés. A partir de celui-ci est défini un score pondéré (B5) qui sera ensuite impacté par son voisinage (B6) pour obtenir un score final (B7). La branche parallèle (B2) est utilisée pour calculer cette diffusion à partir du voisinage des candidats. Il faut alors calculer de la même manière les scores pour les voisins des candidats (B3',B4',B5') afin de finaliser la diffusion (B6). Une fois la phase B terminée, les scores finaux sont transmis à la phase C.*

peut ne jamais le sélectionner (puisque ses voisins peuvent ne jamais être sélectionnés). Afin d'éviter une situation où ces extrema locaux sont isolés, il est possible d'effectuer une diffusion du score d'intérêt.

La solution consiste pour chaque sommet intéressant à diffuser une partie de son intérêt à son voisinage. Pour ce faire, les utilisateurs sélectionnent un degré de diffusion *dif*. Plus *dif* est élevé, plus le diamètre du réseau extrait peut être large (si *dif* est nul, alors le mécanisme de diffusion n'est pas utilisé). Pour faire une diffusion de degré *dif* d'un sommet, il faut calculer le score des sommets non candidats à distance *dif* de celui-ci (B2). Ensuite, une fois le score calculé pour tous les sommets requis (B3',B4',B5'), chaque sommet gagne un pourcentage du $wScore$ du sommet le plus intéressant (i.e. le sommet avec le score le plus élevé) à distance *dif* ou moins (B6). Ce pourcentage est également déterminé par l'utilisateur. Plus il est élevé, plus les scores des sommets vont s'homogénéiser.

Ce mécanisme optionnel peut améliorer la pertinence des sous-réseaux d'intérêt. Cependant, un degré élevé de diffusion peut impacter négativement les performances du processus si le réseau est très connecté. De même, un pourcentage élevé de diffusion homogénéise les scores, ce qui diminue l'impact sur la sélection des scores d'intérêts et des classements qu'ils génèrent. Ce mécanisme doit donc être utilisé avec prudence. Par ailleurs, le fait qu'un sommet théoriquement intéressant soit isolé par son entourage peu intéressant peut ne pas déranger l'expert si celui-ci met l'accent dans son analyse sur les liens ou les communautés. Auquel cas, il n'est pas nécessaire de procéder à une étape de diffusion.

4.1.3 Extraction du sous-réseau

Cette phase commence par le calcul de la liste des sommets choisis (Fig. 4.6, C3). Cette liste contient les sommets sélectionnés dans la liste des candidats pour composer le nouveau sous-réseau. Les sommets sélectionnés par l'utilisateur sont automatiquement dans la liste des sommets choisis. La finalité du processus complet est donc de remplir la liste des sommets choisis en fonction des scores obtenus afin d'avoir un sous-réseau d'intérêt optimal pour l'utilisateur.

Un classement des sommets est effectué en utilisant les scores finaux des phases précédentes (C1) et le sommet ayant le score le plus élevé est ajouté à la liste des sommets choisis (C2). Si le nombre de sommets présent dans la liste des sommets choisis correspond au nombre souhaité par l'utilisateur (C4), on extrait le sous-réseau composé des sommets de la liste et des liens présents entre ces sommets dans le réseau initial (C6).

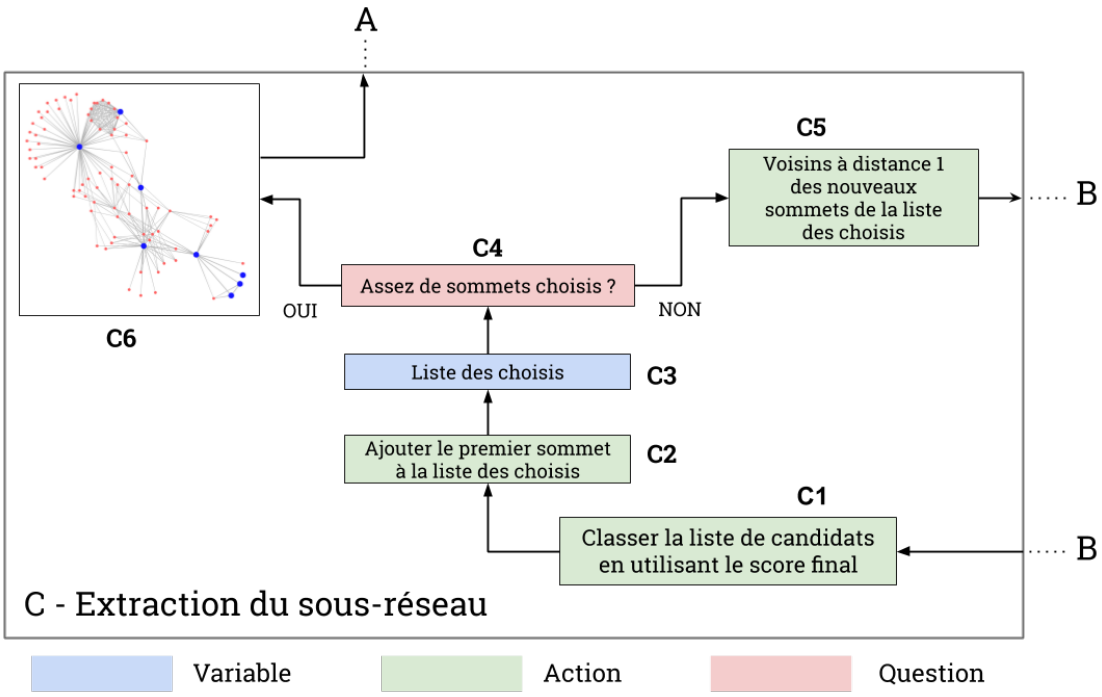


FIGURE *4.6* – ***Phase d'extraction du sous-réseau*** *: La phase C commence en effectuant un classement des sommets à partir de leurs scores finaux (C1). Le sommet en première place est ajouté à la liste des sommets choisis (C2). Cette liste (C3) correspond aux sommets qui seront présentés à l'utilisateur dans le prochain sous-réseau. La quantité de sommets présents dans la liste des choisis est évaluée (C4). Si elle ne contient pas assez de sommets par rapport au nombre souhaité par l'utilisateur, le voisinage du nouveau sommet choisi est ajouté à la liste des candidats (C5) et une nouvelle phase de calcul de l'intérêt commence (B) sur cette nouvelle liste. Si assez de sommets sont présents (C4), un sous-réseau composé des sommets de la liste des choisis et des liens entre eux est extrait et présenté à l'utilisateur (C6). A partir de ce sous-réseau, l'utilisateur pourra sélectionner de nouveaux sommets et ainsi recommencer la procédure à partir de la phase A.*

Si le nombre de sommets n'est pas suffisant, les sommets voisins du dernier sommet choisi sont ajoutés, s'ils n'y sont pas déjà (C5), à la liste des candidats. La mise à jour de la liste de candidats requiert alors que les nouveaux sommets n'ayant pas encore de scores soient traités. La procédure est alors répétée depuis B1. Les scores des sommets précédemment traités n'ont néanmoins donc besoin d'être ré-évalués car leurs scores ne subiront aucun changement.

4.1.4 Processus de génération complet

Une fois ces phases complétées (Fig. 4.7), un sous-réseau est généré. L'utilisateur peut alors sélectionner de nouveaux sommets qui lui semblent pertinents dans ce sous-réseau pour réitérer le processus depuis le début. Le processus complet recommence alors entièrement avec un nouvel ensemble focus enrichi des nouvelles sélections et dé-sélections de l'utilisateur (A2). De nouveaux scores sont calculés engendrant ainsi un nouveau sous-réseau sémantiquement plus proche des nouvelles indications de l'utilisateur. Cette procédure est réitérée jusqu'à satisfaction de l'utilisateur vis à vis des sous-réseaux obtenus.

La répétition de ces étapes permet de générer des sous-réseaux se succédant. Il est néanmoins nécessaire de rappeler que chaque sous-réseau est extrait du graphe initial et généré à partir des sélections utilisateurs provenant d'un sous-réseau père. Chaque sous-réseau ayant alors un sous-réseau père mais potentiellement plusieurs sous-réseaux fils, cette succession n'est pas linéaire et il est alors nécessaire de proposer un moyen de représenter et d'utiliser cette arborescence de sous-réseaux.

4.2 Arbre de traces

La méthodologie de travail de nos experts est éminemment arborescente : lorsqu'une piste d'analyse est suivie, il est fréquent que de nouvelles opportunités apparaissent et ouvrent la voie à de nouvelles pistes à explorer. La méthode `M-QuBE`[3] suit cette méthodologie en proposant à l'expert d'utiliser un "arbre de traces" interactif, inspiré par la notion d'*"history tree"* [14, 71].

Chaque action effectuée par l'utilisateur (sélection/dé-sélection d'un sommet, augmentation/diminution du nombre de sommets requis par sous-réseau ou changement de l'algorithme de positionnement des sommets) génère une "trace" dans l'arbre i.e. un sommet représentant le sous-réseau résultant de cette action. Ce sommet est lié à son père (le sous-réseau sur lequel on a effectué l'action) par une arête dont le type représente l'action effectuée (sélection, augmentation, positionnement).

Cet arbre ne se contente pas d'être descriptif mais propose à l'utilisateur un système de navigation entre tous les sous-réseaux obtenus. Lorsque l'utilisateur clique sur un sommet de l'arbre de trace, le sous-réseau correspondant est sélectionné et affiché. Toute nouvelle action de la part de l'utilisateur va alors générer un sous-réseau dont le père est le sous-réseau sélectionné, permettant ainsi la création d'une nouvelle branche sur l'arbre de trace (Fig. 4.8).

Ce mécanisme fait ainsi écho à la méthode de travail des experts en permettant

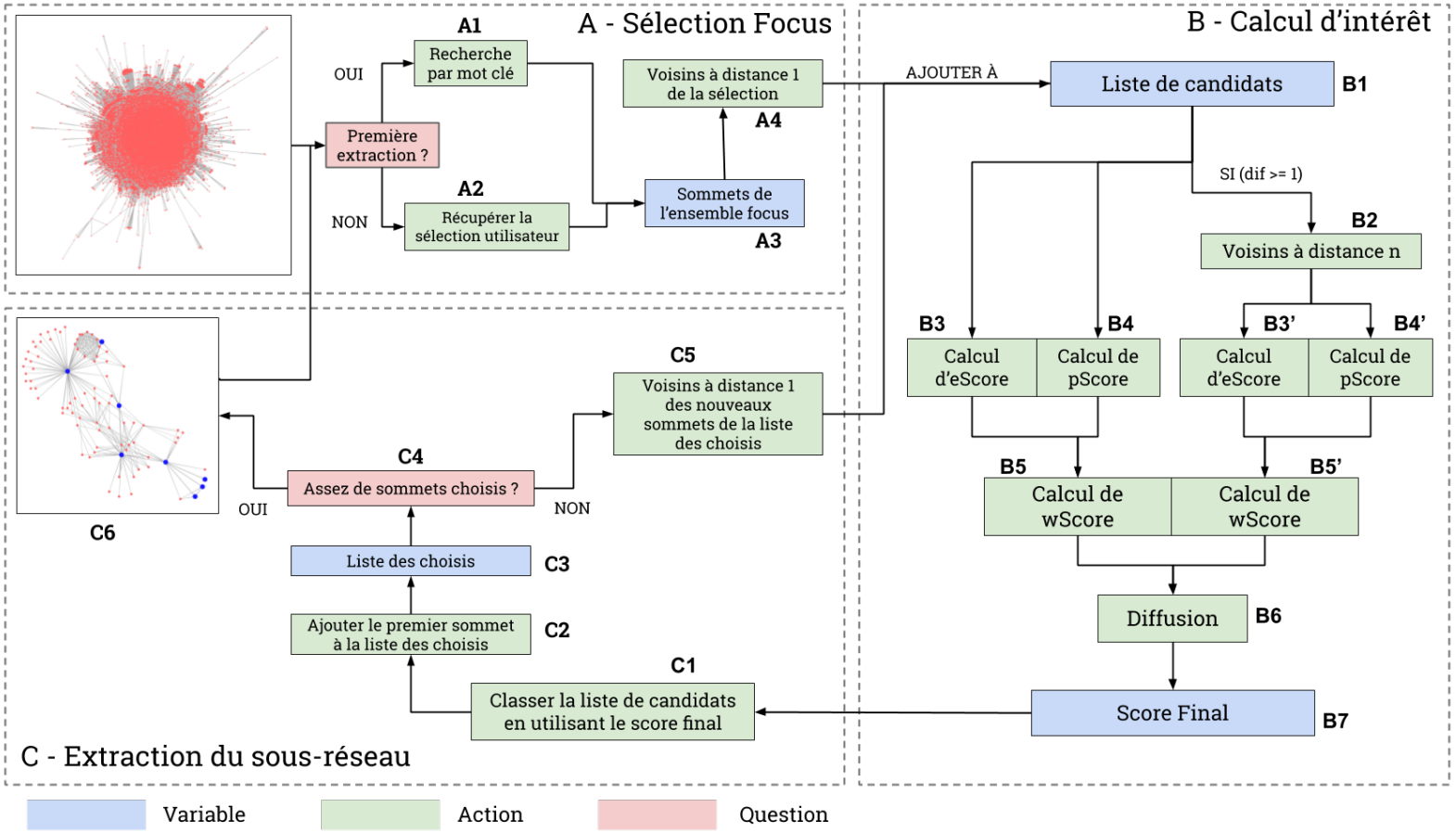


FIGURE *4.7* – ***Algorithme complet de*** `M-QuBE`^{3} *: Schéma non simplifié des interactions entre les différents phases du processus. Le processus se comporte comme un algorithme glouton, une boucle apparaît entre les phases B et C (de C5 vers B1) pour étendre l'évaluation des sommets au voisinage du sommet choisi (C2). Un seconde boucle apparaît entre la phase C et la phase A (de C6 vers A2) lorsque le processus* `M-QuBE`^{3} *est re-sollicité afin d'ajouter un nouveau sous-réseau à la succession de sous-réseaux déjà générés.*

de suivre et développer plusieurs pistes simultanément et a donc été particulièrement bien accueilli par les experts des données. Pour rendre cela possible sans pertes de performance et sans duplication de données entre les sous-réseaux, nous utilisons la structure de données Tulip [3] basée sur une hiérarchie de graphes. Ceci s'ancre dans la continuité des travaux commencés avec Porgy [66] dont le fonctionnement est aussi basé sur un graphe de graphes assuré par le modèle de données Tulip. Des informations supplémentaires quant à l'implémentation, l'emploi et les retours utilisateurs sont disponibles dans le chapitre 5.

4.3 Synthèse

La méthode M-QuBE^3 a pour objectif de répondre à plusieurs particularités des sciences humaines et sociales. Parce que les sciences humaines et sociales opèrent à une échelle proche des individus et groupes d'individus, M-QuBE^3 propose d'utiliser des vues partielles pour explorer le réseau global : en offrant en permanence des sous-réseaux facilement lisibles et analysables, chaque individu du réseau est accessible et peut servir à étendre l'exploration en offrant de nouveaux sous-réseaux construits autour de ses spécificités. Par ailleurs, tant par la création successive de sous-réseaux que les possibilités offertes par l'arbre de trace, M-QuBE^3 respecte la méthode de travail des experts des sciences humaines et sociales en proposant d'explorer différentes pistes simultanément, de revenir en arrière dans leurs recherches à tout moment et d'approfondir autant que souhaité les pistes qui semblent prometteuses.

Cependant, pour répondre aux trois points précédemment cités en introduction de ce chapitre, il reste encore à prendre en considération le caractère multi-couche inhérent aux réseaux des sciences humaines et sociales. Nous avons vu dans la section 4.1.2 que M-QuBE^3 est construit autour de différents calculs de scores dont l'eScore, un score d'intérêt orienté vers la sémantique du réseau. Dans le chapitre suivant, nous expliquons comment, par l'intermédiaire de ce score au centre du fonctionnement de M-QuBE^3, nous offrons la possibilité aux experts de ces réseaux de pouvoir orienter leurs explorations et recherches en prenant en compte l'importance sémantique des réseaux multi-couches.

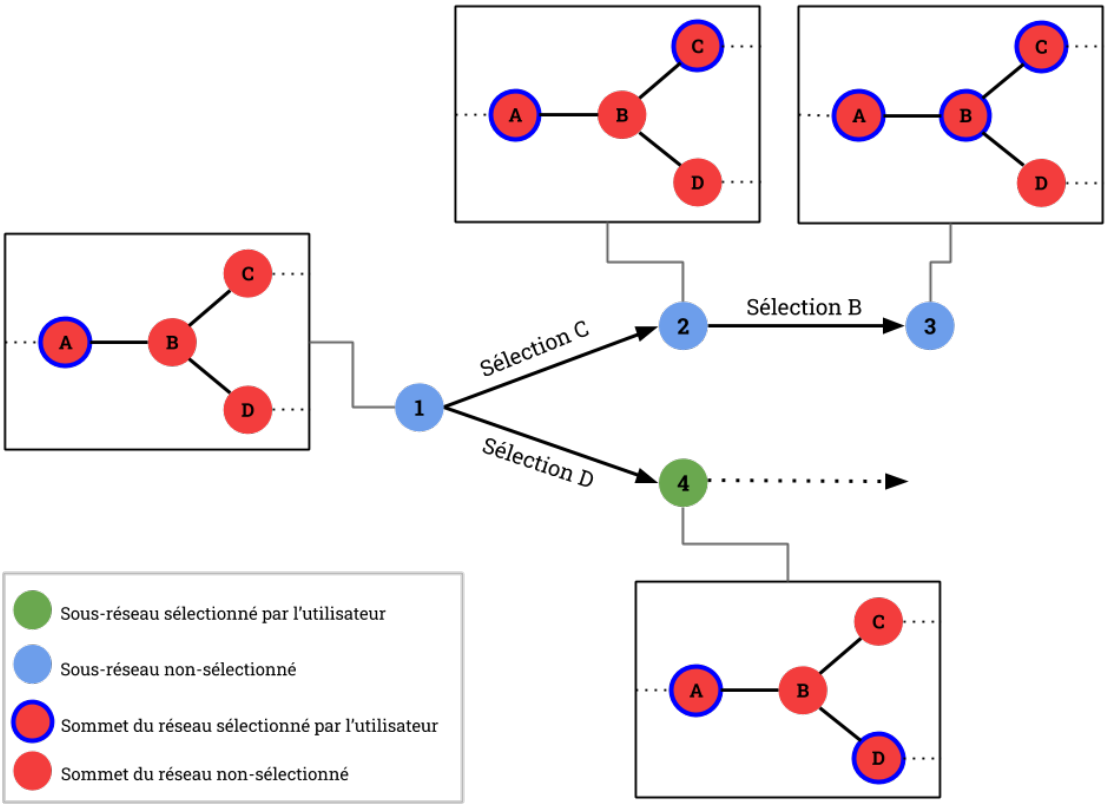


FIGURE *4.8* – ***Exemple d'arbre de traces*** *: Les sommets A,B,C,D représentent des éléments du réseau sur lequel* `M-QuBE`^{3} *est appliqué. Lorsqu'une action de l'utilisateur génère un nouveau sous-réseau, un sommet représentant ce sous-réseau est créé (les sommets 1,2,3,4). Ainsi le sommet 1 de l'arbre de trace représente le premier sous-réseau généré où seul le sommet A a été sélectionné par l'utilisateur (sommet sur-ligné en bleu). Par la suite, lorsque l'utilisateur a sélectionné le sommet C afin de générer un nouveau sous-réseau, le sommet 2 le représentant a été créé. L'arête entre son père (1) et lui indique l'action l'ayant généré. Le sommet vert représente le sous-réseau actuellement sélectionné et visualisé : une fois les réseaux 1,2,3 générés, l'utilisateur a cliqué sur le sommet 1 afin de le sélectionner et de créer une nouvelle branche en sélectionnant D au lieu de C. La sélection est maintenant sur son fils, le sommet 4, qui sera le père de la prochaine action effectuée par l'utilisateur.*

Chapitre 5

Estimation d'intérêt dans les réseaux multi-couches

Sommaire

Quantifier l'intérêt des éléments des structures de données n'est pas nouveau. De nombreux auteurs ont établi des métriques d'estimation d'intérêt à des fins de visualisation ou de navigation dans des contextes applicatifs précis. Pour les graphes multi-couches, le concept est encore plus pertinent tant cela entre en synergie avec la forte richesse sémantique de ces objets. Le concept s'est néanmoins développé progressivement avec en premier lieu des estimations sur des objets simples en ne prenant en compte que les informations directement accessibles, la topologie de la structure, avant de s'étendre à la sémantique.

Nous commençons donc par présenter les différents éléments marquants de la quantification de l'intérêt (Section 5.1) puis notre version spécialisée pour les réseaux multi-couches (Section 5.2).

5.1 Estimation de l'intérêt

L'estimation d'intérêt est aujourd'hui omniprésente dans énormément de domaines et fonctionnalités que nous utilisons chaque jour. De la recommandation de vidéos [20] aux moteurs de recherches [7, 19] en passant par l'"Eye-Tracking" [72], il est devenu capital dans la masse de données accessibles de pouvoir isoler ce qui est pertinent pour l'utilisateur. Pour répondre à ce problème, une solution fréquemment utilisée est d'établir un classement des éléments accessibles (comme les exemples précédents et ceux à venir). Il est alors nécessaire d'avoir un critère permettant de quantifier l'intérêt que l'utilisateur a de chaque élément afin de pouvoir les hiérarchiser.

Si la problématique est très actuelle (plateformes de vidéos à la demande, moteurs de recherche, streaming audio et vidéo, etc.), le concept de score d'intérêt ne l'est pas. Déjà en 1986, les travaux de Furnas [27] proposent une première approche de quantification de l'intérêt appelée $DOI_{fisheye}$ (*Degree of Interest Fisheye*) S'appliquant sur les arbres. Le $DOI_{fisheye}$ est alors utilisé afin de simplifier visuellement des fichiers textes hiérarchisés comme du code C en n'affichant que les lignes ayant un score d'intérêt suffisant. Pour ce faire, chaque ligne du texte correspond à un élément de l'arbre et la profondeur de la ligne dans la hiérarchie des blocs de code est équivalente à la profondeur dans l'arbre à partir de la racine.

Pour chaque sommet x de l'arbre (i.e. chaque ligne du code), le $DOI_{fisheye}$ est calculé ainsi :

$$DOI_{fisheye}(x|y) = API(x) - D(x, y)$$

où y est un sommet de référence appelé le sommet $focus$ représentant le point d'attention actuel de l'utilisateur (ici la ligne actuellement éditée ou étudiée), où $API(x)$ ("à priori") est une fonction qui représente l'intérêt "absolu" d'un élément de l'arborescence (indépendemment de y) pour l'utilisateur et où $D(x, y)$ est une fonction qui correspond à la distance entre les sommet x et y dans l'arbre. Dans cet exemple, API correspond à la distance du sommet x par rapport à la racine de l'arbre. La racine représentant le plus haut niveau de l'arborescence du code (le bloc de la fonction elle même), plus une ligne est intriquée dans des blocs moins son score est élevé. Ainsi, les instructions profondément enfouies dans des blocs de codes, eux même enfouis dans d'autres blocs, etc. sont jugées moins pertinentes pour l'utilisateur et seront donc moins susceptibles d'être montrées à l'utilisateur. Enfin, la fonction D a pour effet de renforcer le score final des éléments à proximité de l'élément focus y. Ainsi, plus un élément va être éloigné du focus, moins il sera

jugé intéressant. Avec ces deux fonctions, le $DOI_{fisheye}$ s'appuie sur la topologie de l'arbre afin de supposer une vue simplifiée plus intéressante pour l'utilisateur. Celui-ci fixe ensuite un seuil et toutes les lignes ne dépassant pas ce seuil sont agrégées et remplacées par un "..." dans la marge du texte visualisé final.

Si le $DOI_{fisheye}$ de Furnas concerne exclusivement les arbres, le concept a ensuite été étendu (comme pour le DOI Tree [16] interactif) ou appliqué à divers objets (ontologies [36], coordonnées parallèles [17], etc.) par d'autres auteurs. Parmi ces travaux, certains mettent l'accent sur la prise en compte de la sémantique des données en utilisant des variations du $DOI_{fisheye}$ [38,77] dont Van Ham et Perer ont ensuite proposé une généralisation appliquée aux graphes [76]. Dans cette généralisation, ils proposent une nouvelle version enrichie du DOI permettant d'extraire un sous-graphe pertinent pour un utilisateur à partir des informations topologiques et sémantiques d'un ensemble sinon difficile à analyser.

En plus de la distance et de la fonction API déterminées à partir des informations structurelles du graphe, Van Ham et Perer proposent d'utiliser une fonction UI (*User Interest*) basée sur la sémantique du graphe (mots-clés, tags, valeurs d'attributs, etc.). Ce DOI est ainsi défini :

$$DOI(x|y,z) = \alpha.API(x) + \beta.UI(x,z) + \gamma.D(x,y)$$

où UI utilise une fonction utilisateur z et où α, β et γ sont des constantes servant de levier afin de moduler l'importance des informations structurelles (API, D) ou sémantiques (UI) dans l'estimation de l'intérêt. La fonction z est une requête de l'utilisateur représentant son intérêt pour les informations sémantiques issues du sommet x. Plus z tend à être validée par les informations issues du sommet x plus le score de UI est élevé. Par exemple, il peut s'agir de la similarité d'un mot clé z avec un attribut texte du sommet x, une note par rapport à un attribut numérique, etc. L'essentiel pour une utilisation efficace du DOI est donc de faire correspondre au mieux les différentes fonctions aux besoins et objectifs des utilisateurs, ce qui peut être fait en utilisant les diverses métriques et indicateurs utilisés en sciences humaines et sociales (comme les exemples donnés dans les travaux de Mainas vus précédemment [54]).

Cependant, le DOI de Van Ham et Perer peut ne pas convenir dans le cas de graphes spécifiques à cause de sa généricité. C'est par exemple le cas des graphes dynamiques [1] dont une adaptation a été proposée afin de pouvoir bénéficier des spécificités de ces objets. De manière analogue, il est nécessaire d'adapter le calcul d'estimation de l'intérêt pour convenir aux réseaux multi-couches et à `M-QuBE`^{3} en prenant en compte à la fois les différents types de sommets (donc les couches) et

à la fois le caractère itératif de la méthode M-QuBE^3. C'est ce que nous proposons avec notre calcul d'intérêt spécialisé, l'eScore.

5.2 eScore, une métrique adaptée et appliquée aux réseaux multi-couches

L'eScore se propose comme une adaptation itérativement applicable pour les réseaux multi-couches inspirée du $DOI_{fisheye}$ de Furnas [27] et de la dimension sémantique de Van Ham et Perer [76]. Pour ce faire, un score est également calculé individuellement pour chacun des sommets du réseau mais en prenant cette fois en compte les informations sémantiques issues des couches et un ensemble de sommets focus évoluant en fonction des sélections de l'utilisateur.

Parce que M-QuBE^3 se veut interactif, eScore doit aussi se baser sur les choix et actions de l'utilisateur. Celui-ci va donc intervenir de deux manières afin d'impacter le calcul d'intérêt : dans un premier temps, et préalablement à toute procédure, à travers le choix d'un ensemble de contraintes et d'objectifs liés aux couches du réseau (les "fonctions de pilotage") et, dans un second temps, en exploitant les sommets jugés pertinents sélectionnés par l'utilisateur qui définissent l'"ensemble focus" (voir sous-section 4.1.1).

Dans M-QuBE^3, chaque nouvelle itération génère un nouveau sous-réseau à partir d'une nouvelle sélection. L'eScore va alors se comporter de la même manière en évoluant en fonction de chaque nouvelle sélection. L'utilisateur se voit ainsi proposer un nouveau sous-réseau donc le calcul a été impacté par eScore et donc la sélection d'un de ses sommets va impacter le prochain calcul d'eScore, impactant le prochain sous-réseau, etc.

A noter que la sélection de l'ensemble focus dans M-QuBE^3 s'effectue en amont du calcul de l'eScore (respectivement phase A et B dans la Fig. 4.3). Cette sélection s'effectue par recherche par mots-clés ou via une visualisation interactive. L'eScore calculant une estimation d'intérêt en fonction d'une sélection utilisateur changeante, il est alors possible d'utiliser l'eScore dans n'importe quel processus évolutif où le calcul de l'eScore peut être ré-itéré au fur et à mesure des nouveaux éléments pré-sélectionnés par l'utilisateur.

Dans la suite, nous considérons que l'eScore est utilisé conjointement avec M-QuBE^3. Dans un premier temps, nous définissions formellement l'eScore (Section 5.2.1) afin, dans un second temps, d'expliquer et définir ce que sont les fonctions de pilotage qu'il utilise (Section 5.2.2).

5.2.1 Formalisation

En s'inspirant du modèle de Kivelä *et al.* [41], notre réseau multi-couches est défini par $G(V, L, E)$ où V est l'ensemble des sommets, L l'ensemble des couches tel que $\forall l \in L, l : (0,1)^{|V|}$ et E l'ensemble des arêtes tel que $E : (V, L) \times (V, L)$. Pour chaque $v \in V$, $b_l(v) : (0,1)^{|L|}$ renvoie un vecteur binaire indiquant les couches auxquelles appartient le sommet v.

La volonté utilisateur pour chaque sommet v de V est définie par un ensemble F de fonctions. Chaque f de F s'applique à un sous-ensemble de couches $L' \subseteq L$ (aspects) et peut être définie par un vecteur binaire b_{lf} indiquant les couches sur lesquelles s'applique la fonction f tel que : $b_{lf}(f) : (0,1)^{|L|}$. Chaque fonction de F renvoie un score normalisé entre 0 et 1.

L'eScore pour un sommet x compte tenu de l'ensemble focus Y peut être défini ainsi :

$$eScore(x|Y) = \frac{\sum_{i=1}^{|F|} f_i(x, Y, L'_i, b_l(x))}{|F|}$$

où $b_l(x) \subseteq L'_i \subseteq L$.

Le rôle de chaque fonction est de guider la navigation en prenant en compte les différences sémantiques et les différences d'intérêt utilisateur entre les couches. Ces fonctions s'appellent des "fonctions de pilotage". Par exemple, soient deux couches données composées d'un ensemble de sommets représentant des fichiers vidéos dans une des couches et de fichiers audio dans l'autre. Une fonction commune au deux couches peut ne pas avoir de sens. Les différents attributs des couches peuvent contraindre à une différenciation lors de l'estimation de l'intérêt et ainsi contraindre à déterminer des méthodes différentes. Il est aussi possible que l'utilisateur ne porte pas une attention égale aux différentes couches et souhaite focaliser son attention davantage sur les vidéos que sur les pistes audio. L'eScore propose donc d'établir une attribution des fonctions de pilotage en fonction des différentes associations de couches possibles.

Dans notre exemple, il est possible d'avoir une fonction attribuée à chacune des couches audio et vidéo et une fonction commune aux deux couches. Ainsi un sommet représentant un fichier vidéo sera considéré par une fonction de pilotage spécialisée qui prendra en compte son type et ses attributs spécifiques et sera aussi considéré par la fonction de pilotage globale pour permettre une comparaison et une analyse plus générale.

Autre exemple plus formel pour un réseau à 4 couches : une fonction attribuée

à l'association de couches $b_{lf}(f) = (0, 1, 1, 0)$ va influer sur les scores de chaque sommet x dont, pour $\forall n \in \{0, 1\}$, $b_l(x) = (n, 1, n, n)$ ou $(n, n, 1, n)$ (Fig.5.1).

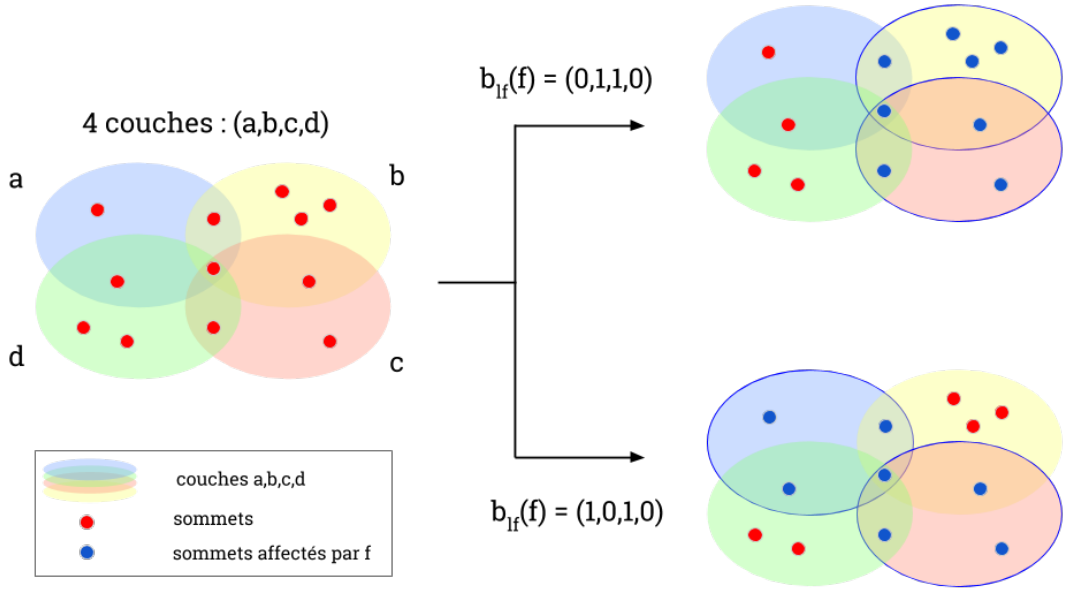


FIGURE *5.1* – ***Domaine d'application des fonctions de pilotage*** *: Le vecteur binaire lié à une fonction de pilotage représente son domaine d'application i.e les différentes couches sur lesquelles elle s'applique. Un sommet est donc utilisé par chaque fonction comprenant dans son domaine d'application la couche à laquelle il appartient.*

Actuellement, les fonctions de pilotage doivent être sélectionnées et paramétrées avec les experts des données pour s'assurer qu'elles soient en accord le plus étroitement possible avec leur volonté afin de s'assurer d'une pertinence optimale. Dans la partie suivante, nous allons proposer une catégorisation de ces fonctions et quelques exemples ayant été utilisés.

5.2.2 Fonctions de pilotage

Déterminer une mesure d'intérêt à partir d'un sommet peut découler de la sémantique des données (recherche de mots dans des champs texte, sélection manuelle de sommets par l'utilisateur, analyse des attributs des sommets), de la topologie du réseau (i.e centralité d'intermédiarité, degrés, faire partie d'une clique ou

d'une triade, etc.) ou, parfois, des deux simultanément. Les fonctions de pilotage suivent ce schéma en pouvant s'orienter à la fois vers la sémantique et la topologie.

En plus de cela, nous distinguons ces fonctions selon deux critères : l'"interaction" (dépendant de l'ensemble focus ou constant) et le "domaine d'application" (couche simple ou association de couches). Les fonctions de pilotage sont définies à travers ces deux critères en fonction de leurs rôles et objectifs afin d'obtenir une formalisation mathématique de la volonté utilisateur (Fig. 5.2).

Interaction	Application	Exemples
Sélection Utilisateur	Couche simple	Nombre défini de politiciens souhaité dans la sélection utilisateur (un sommet appartenant à la couche politicien est favorisé si la quantité choisie n'est pas atteinte)
	Association de couches	Homogénéité du nombre de sommets sélectionnés par l'utilisateur à travers les différentes couches reliés à des types de documents (article, vidéo, interview, etc.)
Statique	Couche simple	Connectivité d'une personne dans la couche personne (les sommets des autres couches ne sont pas considérés)
	Association de couches	Calcul de centralités sur l'intégralité du réseau

FIGURE 5.2 – ***Classification des fonctions de pilotage*** *: Chaque fonction de pilotage peut être classée selon deux critères. Elle peut être dépendante de la sélection utilisateur ou être constante/statique. Une fonction basée sur la sélection utilisateur rendra le calcul d'intérêt dépendant de l'action de l'utilisateur et peut rendre ainsi le processus utilisant le score interactif. En plus de cela, dans chacune de ses catégories, elle peut être liée à un couche simple ou à un ensemble/association de couche. Une association de couche peut comprendre l'ensemble de toutes les couches du réseau si la fonction doit s'appliquer sur tous les sommets du réseau.*

Interaction

L'interaction d'une fonction de navigation se définit par l'impact donné aux actions de l'utilisateur sur le calcul de score entre les différentes exécutions de `M-QuBE`^{3}/eScore. Une fonction peut appartenir à deux catégories : soit basée sur la sélection utilisateur (tous les sommets de l'ensemble focus) soit statique (indépendant de tout choix utilisateur).

Fonctions basées sur la sélection utilisateur Ces fonctions sont basées sur la sélection utilisateur. Parce qu'elles utilisent l'ensemble focus, le résultat peut varier d'une itération à l'autre.

Leur objectif est d'une part de renforcer l'interactivité en mettant l'utilisateur aux commandes et d'autre part de permettre à la méthode de s'adapter si des contraintes ou des objectifs doivent être respectés lors de la recherche.

Un exemple est l'homogénéité de type dans l'ensemble focus. Les utilisateurs veulent avoir un nombre équivalent de sommets des différents types possibles dans leur sélection. Il est donc nécessaire de maximiser les scores des sommets qui améliorent l'homogénéité de cette sélection s'ils doivent être sélectionnés (et inversement minimiser le score des sommets déséquilibrant davantage l'homogénéité de la sélection s'ils sont sélectionnés). Cette procédure est similaire à un calcul d'optimisation d'entropie. En uniformisant le nombre de chaque type dans la sélection, les scores des sommets deviennent alors égaux car leurs sélections seraient d'un impact équivalent sur l'homogénéité. L'utilisateur ayant ainsi un nombre maximum de choix pour sa sélection, l'entropie est alors maximisée. Un cas pratique utilisant cet exemple est présenté dans le chapitre 6.

Fonctions constantes Une fonction constante est une fonction avec aucun prérequis de la part de l'utilisateur pour calculer son score. Il est néanmoins possible de l'appliquer sur une ou plusieurs couches (voir paragraphe suivant). Ces fonctions peuvent être topologiques ou sémantiques. Par exemple, nous calculons dans notre réseau un classement basé sur les degrés de tous les sommets (et donc de la topologie du réseau) ou différents types de centralité. Pour une fonction orientée vers la sémantique, on peut utiliser par exemple un score de proximité entre un mot clé et des attributs des sommets, comme utilisé par Van Ham et Perer [76].

Parce que ces fonctions sont indépendantes du contexte, elles peuvent être calculées antérieurement à toute action utilisateur et leurs résultats peuvent être conservés entre les différentes itérations en cas de processus avec de multiples calculs de l'eScore.

Domaine d'application

Le domaine d'application d'une fonction de pilotage correspond aux couches sur laquelle elle va avoir une influence. Les fonction de pilotage peuvent ainsi soit s'appliquer sur une couche soit sur une association de couches. Il est à noter que, comme dit précédemment, des chevauchements sont possibles entre les différentes

fonctions. Ainsi, une couche donnée peut être comprise et concernée par plusieurs fonctions mono-couches et/ou plusieurs fonctions d'association de couches.

Couche simple Il est parfois nécessaire de pouvoir définir un objectif spécifique pour une catégorie de sommets du réseau. Dans notre exemple, les historiens voulaient trouver des personnes importantes liées à autant d'autres personnalités que possible dans le réseau. Nous avons donc ajouté un calcul de degré interne à la couche personne (en ne considérant que les liens entre deux sommets appartenant à la couche personne). De telles fonctions peuvent aussi être utilisées pour simplement pondérer une couche en particulier du réseau. Si les experts ne sont pas intéressés par un pan des données, il est alors possible d'attribuer uniquement par cet intermédiaire un score bas pour la couche qui est jugée non pertinente.

Association de couches Les fonctions basées sur l'association de couches permettent de mettre en valeur l'interaction entre les différentes couches du réseau ou de faire l'union de certaines couches autour d'un objectif commun. Un exemple orienté sur la topologie est de calculer la centralité dans un sous-réseau composé de sommets inclus uniquement dans une association de couches donnée. Pour la sémantique, l'exemple précédent sur l'homogénéisation de types correspond à nouveau : les types de documents (interview audio, interview vidéo, journal télévisé, article, etc.) sont en réalité une association de couches basée sur les couches interview audio, interview vidéo, etc. sur laquelle s'applique une même fonction d'homogénéisation afin d'obtenir au mieux la même quantité de documents par type dans la sélection. Par ailleurs, il est aussi possible d'instancier une fonction qui s'applique à une association correspondant à toutes les couches du réseau. Ce faisant, les fonctions topologiques classiques (centralité de proximité, centralité d'intermédiarité, degrés,etc.) peuvent être utilisées avec chaque sommet du réseau pour calculer un score.

5.3 Synthèse

Comme vu dans le chapitre précédent, `M-QuBE`^{3} est définie comme une méthode à l'**échelle des individus**, **itérative et arborescente**, en accord avec la méthodologie des experts des sciences humaines et sociales. Parmi nos objectifs initiaux, il reste alors à répondre au caractère multi-couche des réseaux utilisés.

La méthode `M-QuBE`^{3} a été construite autour d'un score d'estimation d'intérêt expressément pour apporter une réponse à cet objectif. En plus de répondre au

caractère multi-couche, il est aussi possible de faire écho aux précédents objectifs en proposant au score de prendre en compte la sélection utilisateur changeante au fil des itérations et de s'adapter en conséquence.

C'est pourquoi nous proposons eScore, une métrique d'estimation de l'intérêt pouvant être utilisée à travers des processus itératifs et spécialisée pour les réseaux multi-couches.

Permettant à la fois d'adapter son traitement à des associations de couches tout en ayant la possibilité de considérer la sélection de l'utilisateur, l'eScore permet ainsi une gestion tant des réseaux multi-couches que des cas plus génériques. Selon la catégorisation que nous avons défini, il est par exemple possible de reproduire l'API de Furnas [27] (en sélectionnant une fonction de pilotage statique appliquée sur l'association de l'ensemble des couches) et l'UI de Van Ham et Perer [76] (en sélectionnant une fonction de pilotage basée sur la sélection de l'utilisateur, ne comportant qu'un seul focus et appliquée à l'association de l'ensemble des couches).

Si eScore peut être utilisé sans `M-QuBE`^{3} et inversement, ces deux travaux ont été conçus et développés afin de fonctionner en synergie pour répondre aux problèmes de nos experts. C'est pourquoi, dans le chapitre suivant, nous détaillons l'implémentation d'une plateforme donnant corps à `M-QuBE`^{3} et eScore afin de valider nos méthodes à travers les cas réels de nos experts des données, issus des sciences humaines et sociales.