Conception de workflows scientifiques fondée sur la structure et la provenance

Résumé

Les systèmes de gestion de workflows scientifiques, (par exemple, Taverna, Kepler, Chimera, Galaxy, Wings) sont de plus en plus utilisés par les scientifiques pour concevoir et exécuter des analyses complexes. Ces analyses sont typiquement centrées sur les données et impliquent d' "enchaîner" les unes après les autres des opérations de récupération de données, de calcul, et des composants de visualisation pour former un pipeline d'analyse exécutable. Un tel pipeline est représenté par un workflow qui est modélisé comme un graphe, où les nœuds représentent les tâches et les arcs dénotent les dépendances d'ordonnancement entre des tâches de calcul. Intuitivement, une spécification de workflow est un cadre pour l'exécution du workflow, qui spécifie l'ensemble des tâches qui sont effectuées et l'ordre à respecter entre les différentes tâches d'exécutions. Selon les données d'entrée et les valeurs de paramètres fournis à la spécification, différentes exécutions (ou runs) de workflow sont obtenus. Un run est alors à son tour représenté par un graphe où chaque sommet représente l'exécution d'une tâche et les arcs sont étiquetés par les données consommées et des produites lors de chaque étape.

Dans cette thèse, nous avons suivi le modèle de workflows présent dans plusieurs systèmes de workflows (en particulier, le système Taverna): nous considérons que les spécifications ont une structure de graphe acyclique dirigé (DAG) et les runs ont les mêmes structures que leurs spécifications. L'objectif principal de nos travaux est de permettre une meilleure réutilisation des workflows en les rendant plus compréhensibles par l'utilisateur.

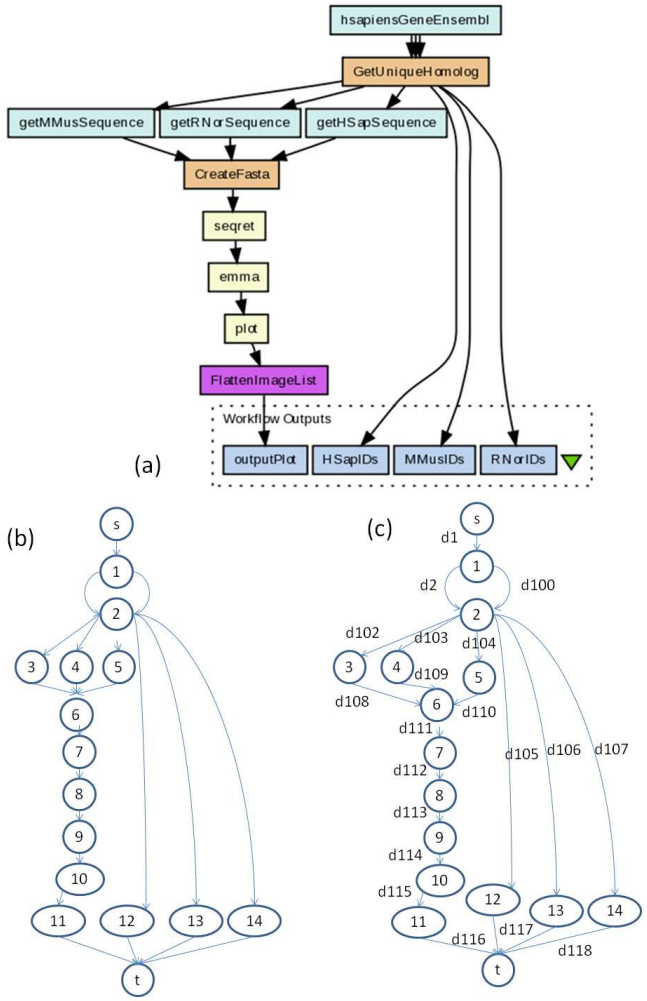

Figure 1: Exemple de workflow issu du système Taverna; (b) représentation sous-forme de graphe; (c) exemple de graphe d'exécution

La Figure 1 propose (a) un exemple de workflow réel issu du système Taverna, (b) sa représentation sous forme de graphe et (c) un exemple de graphe d'exécutionde ce workflow. Tousles graphes (représentant un workflow ou une exécution) ont unesource unique s et un puits unique t, de façon à ce que tout sommet soit sur un chemin de sà t (accessibilité).

Face à la complexité croissante des runs et la nécessité de la reproductibilité des résultats, l'étude de la provenance est devenue un sujet de recherche important.

La provenance (aussi appelée lineage ou pedigree) d'unedonnée produite contient des informations précises sur le processus et les données utilisées pour obtenir cette donnée. Elle est souvent organisée sous forme de graphes de dépendance. La visualisation de ces graphes de dépendance est particulièrement utile pour la réutilisation deworkflows scientifiques, puisque les données, les processus et les dépendances associées à un run de workflow peuvent être clairement compris par les utilisateurs de workflow. En analysant et en créant des visualisations claires de provenance de données, les scientifiques peuvent déboguer leurs analyses et obtenir une meilleure compréhension de leurs résultats. Avec l'aide de la provenance, les scientifiques qui souhaitent effectuer de nouvelles analyses devraient être en mesure de trouver les spécifications de workflowayant des significations identiques ou des intérêtssimilaires qu'ils peuvent réutiliser. Ils peuvent également rechercher des exécutions associées à une spécificationpour comprendre le sens du workflow, ou pour corriger / déboguer une spécification erronée. En outre, les requêtes de provenance structurelles peuvent aider les scientifiques à déterminer pourquoi une donnée a été produite ouàcomprendre comment le processus qui a conduit à créer une donnée a échoué. Par conséquent, les informations de provenance sont clairement utiles pour les systèmes de workflows scientifiques. Toutefois, en raison de la complexité des workflows, les informations de provenance qui sont organisées en un graphe deviennent très grandes, rendant l'exploitation et l'exploration des informations de provenance très complexes pour les utilisateurs. Alors que la plupart des systèmes stockent des données et des dépendances de processus, quelques-uns fournissent des approches efficaces pour accéder aux informations de provenance. En outre, certains systèmes de workflows considèrent des structures de donnée complexes (e.g., listes, arbres $\cdots$), rendant la présentation de la provenance un

point clé.

Dans la dernière décennie, des efforts considérables ont été déployés pour l'amélioration du partage et de la réutilisation des workflows scientifiques. Les objectifs sont alors les suivants: réduire le temps de création de workflow (éviter de "réinventer la roue"); améliorer la qualité à travers le développement desworkflows partagés (en s'appuyant sur l'expertise des utilisateurs précédents) et améliorer la provenance expérimentale au niveau des processus grâce à la réutilisation de workflowsvalidés. Toutefois, comme indiqué par des études récentes, tandis que le nombre de workflows scientifiques disponible augmente avec leur popularité, les workflowsne sont pas (re)utilisés et partagés autant qu'ils pourraient l'être. Il y a plusieurs années, Goderis et al. ontrésumé plusieurs goulots d'étranglement de la réutilisation de workflow et la réorientation, dans laquelle ils affirment que les principales raisons sont les restrictions sur la disponibilité des services, le manque d'un modèle global de la découverte et de la complexité des workflows. Récemment, Zhao et al. affirment que l'un des empêchements principauxà la réutilisation de workflows est liée à la non disponibilité de certaines ressources nécessaires à l'exécution de workflow. Les causes de ces empêchements incluent: le manque de données d'exemple, le manque de connaissance et de documentation sur un environnement d'exécution, les descriptions insuffisantesdes workflows. Plusieurs solutions ont été proposées par Zhao et al.

Dans cette thèse, nous avons mis l'accent en particulier sur le système de gestion de workflowTaverna, qui, dans les dix dernières années, a été populaire au sein de la communauté bioinformatique. Malgré le fait que des centaines de workflows Taverna ont été disponibles pendant des années par le référentiel de workflows publiques myExperiment(http://www.myexperiment.org), leur réutili-

sation par des scientifiques autres que l'auteur original est généralement limitée. Récemment, plusieurs études mettent en évidence la complexité des structures de workflow comme l'une des raisons principales de la réutilisation limitéedes workflows (Taverna) scientifiques. La complexité de la structure de workflow implique le nombre de nœuds et de liens, mais elle est également liéeaux caractéristiques complexes de la structure de workflow. Encore une fois, plusieurs facteurs peuvent expliquer une telle complexité structurelle, y compris le fait que le processus bio-informatique à mettre en œuvre est intrinsèquement complexe, ou le système de workflowsne peut pas fournir l'expressivité nécessaire, forçant les utilisateurs à concevoir des workflows complexes arbitraires. Par conséquent, pour obtenir une structure de workflow plus simple pour un workflow complexe tout en préservant ce qu'il signifie devient particulièrement important.

L'objectif global de cette thèse est d'améliorer la réutilisation des workflows en fournissant des stratégies visant à réduire la complexité des structures de workflow tout en préservant la provenance. Deux stratégies sont introduites.

Tout d'abord, nous proposons une approche de réécriture de la structure du graphe de n'importe quel workflow scientifique (classiquement représentée comme un graphe acyclique orienté (DAG)) dans une structure plus simple, à savoir une structure série-parallèle (SP) tout en préservant la provenance. Les SP-graphes sont simples et bien structurés, ce qui permet de mieux distinguer les principales étapes du workflow. En outre, d'un point de vue plus formel, on peut utiliser des algorithmes polynomiaux pour effectuer des opérations complexes fondées sur les graphes (par exemple, la comparaison de workflows, ce qui est directement lié au problème d'homomorphisme de sous-graphes) lors que les workflows ont des SP-structures alors que ces opérations liées à des problèmes NP-difficiles pour

des graphes qui sont des DAGs sans aucune restriction sur leur structure. Nous avons introduit la notion de préservation de la provenance, conçu l'algorithme de réécriture SPFlow et réalisé l'outil associé.

Deuxièmement, nous proposons une méthodologie avec une technique capable de réduire la redondance présente dans les workflows (en supprimant les occurrences inutiles de tâches). Plus précisément, nous détectons des "anti-modèles", un terme utilisé largement dans le domaine du génie logiciel, pour indiquer l'utilisation des formes idiomatiques qui mène à une conception trop compliquée, et qui doit donc être évitée. Nous avons ainsi conçu l'algorithme DistillFlow qui est capable de transformer un workflow donné en un workflow sémantiquement équivalent "distillé", c'est-à-dire, qui est libre ou partiellement libre des anti-modèles et possède une structure plus concise et plus simple.

Les deux approches principales de cette thèse (à savoir, SPFlow et DistillFlow) sont basées sur un modèle de provenance que nous avons introduit pour représenter la structure de la provenance des exécutions du workflow. La notion de "provenance-équivalence" qui détermine si deux workflows ont une même signification est également au centre de notre travail. Nos solutions ont été testées systématiquement sur de grandes collections de workflows réels, en particulier avec le système Taverna.

Dans le détail, cette thèse est organisée comme suit:

- Le chapitre 1 donne l'introduction de cette thèse en affirmant la motivation, les problèmes de recherche et des contributions.

- Le chapitre 2 présente une collection des notations mathématiques utilisées dans le reste de cette thèse. Sur la base de ces notations, le modèle de

workflow utilisé dans cette thèse est introduit. Nous donnons ensuite une introduction aux graphes série-parallèles et leurs propriétés. A la fin du chapitre 2, nous offrons une introduction aux workflows de système Taverna, le système sur lequel nous avons choisi de travailler principalement.

- Le chapitre 3 commence avec les travaux connexes sur les modèles de provenance, puis propose un modèle pour représenter la provenance des exécutions de workflow. Après, nous donnons une définition de la notion de provenance-équivalence qui peut être utilisé pour déterminer si deux workflows ont une même signification. Enfin, une discussion sur l'extension de notre modèle de provenance pour mieux soutenirdes données des listes est donnée.

- Le chapitre 4 donne d'abord une explication en profondeur de la motivation de la réécriture workflows non-SP en workflows SP. Ensuite, nous introduisons le concept de mesure de la distance de non-SP à SP, qui inspire certaines techniques de transformation de réécriture de graphes non-SP en graphes SP. Nous analysons ensuite les stratégies existantes pour déterminer si elles préservent la provenance et de proposer une nouvelle stratégie provenance-équivalente. Nous introduisons l'algorithme de SPFlow pour transformer graphes non-SP en graphes SP et discutons la complexité et la correction de l'algorithme. Après cela, nous démontrons la faisabilité de notre approche sur les processus scientifiques réels. Nous présentons enfin un outil avec le même nom que notre algorithme qui prend en non-SP Taverna workflow et fournit une version SP du workflow Taverna utilisable.

- Le chapitre 5 donne d'abord une explication approfondie de la deuxième question de la recherche que nous avons considérée en présentant plusieurs

cas d'utilisation. Ensuite, nous introduisons les anti-modèles que nous avons identifiés et les transformations que nous proposons de le faire tout en veillant à ce que la sémantique du workflow reste inchangée. Nous introduisons ensuite l'algorithme de refactorisation DistillFlow. Après cela, nous donnons les résultats obtenus par notre approche sur un grand nombre du véritablesworkflows. Enfin, nous discutons de plusieurs points liés à notre approche.

- Le chapitre 6 présente les conclusions et les futurs travaux.

- L'annexe A présente la mise en œuvre de DistillFlow et l'annexe B présente une étude préliminaire qui cherche pourquoi des workflows scientifiques ont des structures non-SP.

Comme nous proposons deux stratégies de réécrituredes structures desworkflows scientifiques complexes en de plus simples pour les rendre plus faciles à (re) utiliser. Deux séries de contributions sont résumées ci-dessous.

Première série de contributions (ont été publiées dans la 8ème Conférence Internationale IEEE eScience 2012 et les 28èmes Journées de Bases de Données Avancées (BDA) 2012):

- Nous proposons un modèle pour représenter les workflows scientifiques et la provenance des données générées dans leur exécution.

- Nous donnons une définition de la notion de provenance-équivalence qui peut être utilisée pour déterminer si deux workflows produiront toujours les mêmes résultats étant données deux entrées identiques.

- Nous passons en revue plusieurs stratégies de réécriture pour transformer graphes non-SP en graphes SP et nous prouvons qu'ils ne sont pas provenance-équivalent.

- Nous concevons un algorithme provenance-équivalent, nommé "SPFlow", capable de traduire des workflows non-SP en des workflows SP.

- Nous illustrons notre algorithme en fournissant une évaluation de notre approche sur un millier de workflows scientifiques.

- Nous développons un outil basé sur SPFlow, qui prend un workflow Taverna non-SP et fournit une version SP du workflow Taverna utilisable.

Deuxième série de contributions (ont été publiés dans le "BMC Bioinformatics" Journal et le workshop NETTAB 2012 (poster)):

- Nous identifions et détectons automatiquement un ensemble d'anti-modèles qui contribuent à la complexité du workflow structurel.

- Nous concevons une série de transformations de refactorisation pour remplacer chaque anti-modèle par un nouveau modèle sémantiquement-équivalent avec moins de redondance et une structure simplifiée.

- Nous introduisons un algorithme distillant qui prend en un workflow et produit un workflow sémantiquement-équivalente distillé.

- Nous fournissons une implémentation de notre approche de reconstitution que nous évaluons la fois sur les workflows publiques Taverna et sur une collection privée de workflow du projet BioVel.

- Nous développons un outil basé sur DistillFlow, qui prend un workflowTaverna et fournit une nouvelle version du workflow qui est libre ou partiellement libre des anti-modèles.

Nos outils sont disponibles à l'adresse:https://www.lri.fr/∼chenj/.

Mots-clés : workflow scientifique, provenance, provenance-équivalence, graphes séries-parallèles, SP-graphes, Taverna, anti-modèles.