Gestion, analyse et visualisation de graphes d'applications

Contexte de la collaboration

Cette proposition de thèse se place dans le cadre d'une collaboration entre la société CAST (Paris) et le laboratoire LIRIS (Université Lyon 1). La collaboration s'inscrit dans les domaines des Graphes et du Big Data.

Contexte scientifique

Les graphes des applications sont des structures de données extraites automatiquement à partir de l'analyse du code, des fichiers projets (comme les pom.xml dans l'environnement Java), et des structures de données (relationnelles, hiérarchiques ou simples fichiers). CAST Imaging dispose d'une grande base de connaissances de ces graphes qui couvre plus de 50 langages et technologies concernant des applications à la fois modernes exploitant les dernières nouveautés des fournisseurs Cloud comme AWS ou Azure, mais aussi des plus classiques faites autour de JEE, .NET, C, les bases de données relationnelles, etc. Ces applications sont représentées via une interface graphique dédiée par des graphes où les éléments du code (fonction, classes, procédure, tables, fichiers de données, etc.) sont représentés par des nœuds, et les dépendances (appel, héritage, composition, etc ;) entre ces éléments sont représentées par des arêtes/arcs. Par conséquent, l'analyse et la compréhension de ces applications passent naturellement par l'analyse et la compréhension de leurs graphes respectifs.

Objectifs de la thèse

Les graphes des applications peuvent comporter plusieurs millions de nœuds et d'arêtes/arcs. Ils peuvent avoir des représentations lourdes notamment quand on souhaite prendre en compte un maximum d'informations sur les applications. Ils deviennent des multigraphes hétérogènes où les nœuds ne sont pas tous de même nature, les arêtes peuvent décrire plusieurs relations entre une même paire de nœuds, avec des ensembles d'attributs et de poids à la fois sur les nœuds et sur les arêtes. Ces graphes sont riches en informations, mais leur analyse et visualisation dans leurs structures réelles deviennent difficiles. Notre objectif dans cette thèse est donc de s'appuyer sur des modélisations avancées et des algorithmes avancés pour analyser les graphes d'applications et

proposer des représentations simples de ces graphes facilement explorables d'un point de vue algorithmique et compréhensibles d'un point de vue visuel.

Dans un premier temps, nous nous focaliserons sur l'enrichissement des graphes d'applications actuels aux niveaux structurel et sémantique. D'un point de vue structurel, nous identifierons clairement les classes de nœuds (hétérogénéité) et les relations structurelles intra-classes et inter-classes (héritage, inclusion, appels de fonctions, etc.). D'un point de vue sémantique, nous capturerons un maximum d'informations sémantiques sous forme d'attributs, de poids ou de relations entre objets. Nous développerons par la suite des techniques de stockage et d'indexation de ces graphes qui permettraient le passage à l'échelle.

Dans un deuxième temps, nous mènerons une analyse algorithmique des graphes d'applications. Les structures macroscopiques des graphes d'applications sont quelconques, mais l'analyse de leurs sous-graphes, la recherche et la découverte de patterns et des propriétés structurelles permettent une meilleure compréhension du graphe. L'analyse des graphes d'applications s'appuyera à la fois sur des algorithmiques d'exploration de graphes et sur des algorithmes de *machine learning* (clustering, *graph embedding*, etc.). Cette analyse algorithmique des graphes d'applications servira d'une part à mieux comprendre ces graphes et d'autre part à concevoir des représentations simples de ces graphes qui faciliteront leur visualisation. Ces représentations seront des structures résumant le graphe d'application sous forme de structures hiérarchiques multi-niveaux, avec des regroupements et compressions de nœuds/sous-graphes, arêtes, etc. Pour ce faire, nous serons amenés à explorer la littérature des graphes liées aux techniques de décomposition, d'agrégation et de compression de graphes, pour proposer de telles représentations et de les adapter pour qu'elles prennent en compte les contraintes réelles des graphes d'applications (hétérogénéité des nœuds et arêtes, attributs, poids, etc.) et qui préservent au mieux les propriétés structurelles des graphes d'applications.

La visualisation des graphes d'applications est une partie centrale de la thèse. L'objectif est de proposer des méthodes de navigation dans le graphe d'application qui permettent de guider/orienter l'utilisateur dans la découverte et la compréhension du graphe sans le submerger immédiatement avec tout le détail dont nous disposons. D'où l'intérêt des représentations des graphes d'applications décrites dans le paragraphe précédent. Ces représentations simples donneront une visualisation claire qui permettra à l'utilisateur d'effectuer une meilleure analyse visuelle du graphe. Notre objectif dans cette partie visualisation est de développer des algorithmes qui donneront une meilleure performance en temps d'analyse (affichage, exploration etc.), et une visualisation compréhensible (représentations réduites et simplifiées). Pour ce faire, nous explorerons des techniques de visualisation progressive de sorte que l'utilisateur puisse découvrir, à la demande (d'une façon interactive), ou automatiquement, un graphe d'applications et ses représentations pas-à-pas, et de plus, afficher ces parties du graphe de différents angles avec différentes informations en utilisant par exemple des vues 3D des représentations des graphes d'applications.

Finalement, il est à noter que ces graphes d'applications sont dynamiques car les applications sont mises à jour régulièrement. Nous serons amenés à automatiser les mises à jour sur les graphes d'applications.

Localisation
Le/la doctorant.e effectuera sa recherche à la fois dans la société CAST SA et dans l'équipe GOAL du laboratoire LIRIS.

Candidature
Les candidat.e.s ayant obtenu un M2 recherche/ingénieur en informatique, intéressé.e.s, disposant de connaissances approfondies en algorithmique des graphes, *machine learning*, big data et programmation sont prié.e.s d'envoyer leur CV détaillé, une lettre de motivation pour le sujet et des relevés de notes (avec le classement si possible) aux emails suivants, avant le 30 avril 2023 :

- Hamamache Kheddouci : hamamache.kheddouci@liris.cnrs.fr,
- Olivier Bonsignour : o.bonsignour@castsoftware.com,
- Damien Charlemagne : d.charlemagne@castsoftware.com
- Salma Nagbi : s.nagbi@castsoftware.com

Quelques références sur le sujet

1. Giraph (http://giraph.apache.org)
2. GraphLab (https://turi.com)
3. PGX (http://www.oracle. com/technetwork/oracle-labs/parallel-graph-analytics/overview)
4. GraphStream (http://graphstream-project.org),
5. Structure101 https://structure101.com
6. Codesee https://www.codesee.io
7. CodeSonar https://www.grammatech.com/codesonar-cc
8. D. Auber, D. Archambault, R. Bourqui, M. Delest, J. Dubois, B. Pinaud, A. Lambert, P. Mary, M. Mathi- aut, and G. Melancon. Tulip III. In Encyclopedia of Social Network Analysis and Mining, pages 2216–2240. Springer, 2014
9. W. De Nooy, A. Mrvar, and V. Batagelj. Exploratory social network analysis with Pajek, volume 27. Cambridge University Press, 2011.
10. Habib, M. and Paul, C. (2010). A survey of the algorithmic aspects of modular decomposition. *Computer Science Review*, 4(1):41–59.
11. Lagraa, S., Seba, H., Khennoufa, R., M'Baya, A., and Kheddouci, H.. A distance measure for large graphs based on prime graphs. *Pattern Recognition*, 47(9) 2014:2993 – 3005.
12. Yike Liu, Tara Safavi, Abhilash Dighe, Danai Koutra: Graph Summarization Methods and Applications: A Survey. ACM Comput. Surv. 51(3): 62:1-62:34 (2018)
13. Hongyun Cai , Vincent W. Zheng , Kevin Chen-Chuan Chang: A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications. IEEE Trans. Knowl. Data Eng. 30(9): 1616-1637(2018)
14. Victor Lequay, Alexis Ringot, Mohammed Haddad, Brice Effantin, Hamamache Kheddouci: GraphExploiter: Creation, Visualization and Algorithms on graphs. ASONAM 2015: 765-767
15. Shivani Choudhary, Tarun Luthra, Ashima Mittal, Rajat Singh. A Survey of Knowledge Graph Embedding and Their Applications. Arxiv https://arxiv.org/pdf/2107.07842.pdf