2019-01628 - PhD Position F/M Distributed link prediction in large complex graphs: application to biomolecule interactions

Type de contrat : CDD de la fonction publique
Niveau de diplôme exigé : Bac + 5 ou équivalent
Function : Docteurant
Niveau d'expérience souhaité : Jeune diplômé

Contexte et atouts du poste

Today, vast and diverse sources of data exist for almost every scientific domain, making their integration and intelligent exploitation challenging. Indeed, complex data require expressive data representation models such as graph representation. The Linked Open Data (LOD) movement along with the FAW (Findability, Accessibility, Interoperability, Reusability) data principles are intended to facilitate heterogeneous data integration and analyses. In the LOD context, graphs are called knowledge graphs as they encompass domain ontologies for typing objects and describing their relationships. Semantic web languages (RDFS, OWL, SPARQL) have reached an interesting level of maturity on which ambitious machine learning techniques can rely. Interestingly, big data and NoSQL solutions make possible web-scale data analyses. So far, such analyses on dedicated big-data architectures are often limited to MapReduce scenarios on rather simple data models (key-value oriented, homogeneous graphs with only one type of nodes and one type of edges). Graph databases, as one NoSQL approach, allow for rich representation of multi-typed attributed nodes and edges. This better expressivity comes with a cost as graph and program distribution is not an easy task.

The objective of this PhD project is to make progress to the state of the art of link prediction problem in knowledge graphs in a distributed setting ([9],[23]) in the context of predicting drug-target associations. Predicting new targets for known drug molecules (often called drug repositioning [5]) offers the possibility of using existing drug molecules in new ways, which is far cheaper and much less time-consuming than developing new drug molecules from scratch. However, the computational challenge here is to make link prediction algorithms capable to take into account the existence of attributes and labels on both nodes and edges in knowledge graphs. The proposed approaches will be evaluated using web-scale knowledge graphs for inferring missing links (data completion). YAGO, DBpedia, and synthetic benchmarks are usable for such evaluation and validation purposes [4].

Mission confiée

Mission:

This PhD thesis project aims to develop scalable link prediction methods in large and complex graphs. More specifically, the aims of this thesis project are:
- to propose link prediction approaches in knowledge graphs based on both graph topology and neighborhood constraints to be defined;
- to design scalable implementations of the proposed approaches for distributed architectures. In this context, the use of big graph processing frameworks such as Pregel, Trinity, GraphLab and BLADYS need to be studied [6];
- to define evaluation and validation protocols for the proposed algorithms in the context of web-scale knowledge graphs;
- to apply the approach to the prediction of the drug-target associations.

This project will be carried out mainly within the Capsid team at INRIA Nancy which combines expertise in knowledge graphs, distributed graph computing [6] and drug-target interactions (https://capsid.loria.fr). Achieving the objectives of the thesis will involve acquiring knowledge and understanding of the current state of the art in link prediction in large and complex graphs. An important aspect of this project will be to explore the use of big graph processing frameworks in order to design scalable implementations of proposed link prediction methods in knowledge graphs. The proposed techniques will be implemented on a local cluster and evaluated using publicly available data. This project will develop novel and practical link prediction algorithms that will be applied to predict drug targets. This will help to satisfy an important and current research need in drug repositioning. The developed software will be made publicly available.

Required qualification:
Candidates must have a master degree in computer science, mathematics, or one of the physical sciences. Good programming skills in an object-oriented programming language such as JAVA or C++ are essential. Experience of NoSQL solutions (Neo4j, Titan, MongoDB), parallel/distributed programming (Spark, Hadoop, Flink) and graph processing frameworks (Pregel, GraphLab, GraphX) is also desirable but not essential. A strong interest in structural biology would also be highly desirable.

Advantages:
- Duration: 3 years
- Starting date: between Oct. 1st 2019 and Jan. 1st 2020 - Salary : 1 982 euros gross monthly (about 1 676 euros net).
- Medical insurance is included.

Help and benefits:
Possibility of free French courses
Help for accommodations
Help for the resident card procedure and for husband/wife visa
Lunch cost at Inria canteen is about 3 €

References:

Informations générales

- Thème/Domaine : Biologie numérique
- Statistiques (Big data) (BAP E)
- Ville : Villers-lès-Nancy
- Centre Inria : CR Nancy - Grand Est
- Date de prise de fonction souhaitée : 2019-09-01
- Durée de contrat : 3 ans
- Date limite pour postuler : 2019-05-24

Contacts

- Equipe Inria : CAPSID
- Directeur de thèse : Andhi Sabeur / sabeur.andhi@loria.fr

A propos d’Inria

Inria, l’institut national de recherche dédié aux sciences du numérique, promeut l’excellence scientifique et le transfert pour avoir le plus grand impact. Il emploie 2400 personnes. Ses 200 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3000 scientifiques pour relever les défis des sciences informatiques et mathématiques, souvent à l’interface d’autres disciplines. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 160 start-up. L’institut s’efforce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l’économie.

Consignes pour postuler

Sécurité défense :
Ce poste est susceptible d’être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L’autorisation d’accès à une zone est délivrée par le chef d’établissement, après avis ministériel favorable, tel que défini dans l’arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l’annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.

Informations générales

Type de contrat :
CDD de la fonction publique
Niveau de diplôme exigé :
Bac + 5 ou équivalent
Function :
Docteurant
Niveau d'expérience souhaité :
Jeune diplômé

Contexte et atouts du poste

Today, vast and diverse sources of data exist for almost every scientific domain, making their integration and intelligent exploitation challenging. Indeed, complex data require expressive data representation models such as graph representation. The Linked Open Data (LOD) movement along with the FAW (Findability, Accessibility, Interoperability, Reusability) data principles are intended to facilitate heterogeneous data integration and analyses. In the LOD context, graphs are called knowledge graphs as they encompass domain ontologies for typing objects and describing their relationships. Semantic web languages (RDFS, OWL, SPARQL) have reached an interesting level of maturity on which ambitious machine learning techniques can rely. Interestingly, big data and NoSQL solutions make possible web-scale data analyses. So far, such analyses on dedicated big-data architectures are often limited to MapReduce scenarios on rather simple data models (key-value oriented, homogeneous graphs with only one type of nodes and one type of edges). Graph databases, as one NoSQL approach, allow for rich representation of multi-typed attributed nodes and edges. This better expressivity comes with a cost as graph and program distribution is not an easy task.

The objective of this PhD project is to make progress to the state of the art of link prediction problem in knowledge graphs in a distributed setting ([9],[23]) in the context of predicting drug-target associations. Predicting new targets for known drug molecules (often called drug repositioning [5]) offers the possibility of using existing drug molecules in new ways, which is far cheaper and much less time-consuming than developing new drug molecules from scratch. However, the computational challenge here is to make link prediction algorithms capable to take into account the existence of attributes and labels on both nodes and edges in knowledge graphs. The proposed approaches will be evaluated using web-scale knowledge graphs for inferring missing links (data completion). YAGO, DBpedia, and synthetic benchmarks are usable for such evaluation and validation purposes [4].

Mission confiée

Mission:

This PhD thesis project aims to develop scalable link prediction methods in large and complex graphs. More specifically, the aims of this thesis project are:
- to propose link prediction approaches in knowledge graphs based on both graph topology and neighborhood constraints to be defined;
- to design scalable implementations of the proposed approaches for distributed architectures. In this context, the use of big graph processing frameworks such as Pregel, Trinity, GraphLab and BLADYS need to be studied [6];
- to define evaluation and validation protocols for the proposed algorithms in the context of web-scale knowledge graphs;
- to apply the approach to the prediction of the drug-target associations.

This project will be carried out mainly within the Capsid team at INRIA Nancy which combines expertise in knowledge graphs, distributed graph computing [6] and drug-target interactions (https://capsid.loria.fr). Achieving the objectives of the thesis will involve acquiring knowledge and understanding of the current state of the art in link prediction in large and complex graphs. An important aspect of this project will be to explore the use of big graph processing frameworks in order to design scalable implementations of proposed link prediction methods in knowledge graphs. The proposed techniques will be implemented on a local cluster and evaluated using publicly available data. This project will develop novel and practical link prediction algorithms that will be applied to predict drug targets. This will help to satisfy an important and current research need in drug repositioning. The developed software will be made publicly available.

Required qualification:
Candidates must have a master degree in computer science, mathematics, or one of the physical sciences. Good programming skills in an object-oriented programming language such as JAVA or C++ are essential. Experience of NoSQL solutions (Neo4j, Titan, MongoDB), parallel/distributed programming (Spark, Hadoop, Flink) and graph processing frameworks (Pregel, GraphLab, GraphX) is also desirable but not essential. A strong interest in structural biology would also be highly desirable.

Advantages:
- Duration: 3 years
- Starting date: between Oct. 1st 2019 and Jan. 1st 2020 - Salary : 1 982 euros gross monthly (about 1 676 euros net) during the first and the second years. 2 085 euros the last year (about 1 676 euros net).

Help and benefits:
- Medical insurance is included.

References:
Supervision and contact:
Sabeur Aridhi, sabeur.aridhi@loria.fr, https://members.loria.fr/SAridhi
Malika Smail-Tabbone, malika.smail@loria.fr, https://members.loria.fr/MSmail

Principales activités
Thesis tasks:
Study existing link prediction algorithms in homogeneous graphs
Extend the approach to knowledge graphs (by considering node and edge properties)
Evaluate on well-known public RDF datasets
Apply to biological knowledge graphs describing drug-target interactions

Compétences
Candidates must have a master degree in computer science, mathematics, or one of the physical sciences. Good programming skills in an object-oriented programming language such as JAVA or C++ are essential. Experience of NoSQL solutions (Neo4j, Titan, MongoDB), parallel/distributed programming (Spark, Hadoop, Flink) and graph processing frameworks (Pregel, GraphLab, GraphX) is also desirable but not essential. A strong interest in structural biology would also be highly desirable.

Rémunération
Monthly salary after taxes: around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year (medical insurance included).