Titre de la thèse : Décompositions tensorielles couplées pour la détection de phénomènes de diffusion dans les réseaux sociaux

Laboratoire d'accueil

Laboratoire d'Informatique de Bourgogne (LIB EA 7534), Université de Bourgogne, Dijon

Spécialité du doctorat préparé

Informatique

Mots clés

Analyse de données, Décompositions tensorielles, Détection de phénomènes de diffusion

Descriptif du sujet

L'analyse des données vise à extraire de la valeur des données. Toutefois, c'est un processus complexe, qui peut faire appel à des données provenant de différentes sources, ayant différentes caractéristiques, étant stockées dans différents systèmes utilisant différents modèles, etc. De plus, le type d'analyse à effectuer peut imposer des contraintes qui limitent les algorithmes utilisables. Cette thèse s'inscrit dans le projet interdisciplinaire Beelzebot (ANR-23-CE38-0002-01), dont le but est de détecter des armées de robots sur Twitter, tout en fournissant des résultats interprétables par les analystes métier et ainsi produire des alertes lors de campagnes de désinformation. Dans ce contexte, l'utilisation des algorithmes supervisés est difficile. Les comportements des robots changent rapidement pour s'adapter aux méthodes de détection, nécessitant de produire régulièrement de nouveaux jeux d'entraînement pour les algorithmes, alors que l'on constate que les experts ont de plus en plus de mal à distinguer un robot du compte d'un humain [2]. Les informations extraites des interactions et des activités des utilisateurs sur les réseaux sociaux sont de plus en plus utilisées pour détecter les activités des robots en tant que phénomènes de diffusion plutôt qu'en se basant uniquement sur des informations concernant un compte individuel [1].

Les tenseurs sont des objets mathématiques multi-dimensionnels [8], capables d'intégrer les données provenant de différents modèles de données, tels que le modèle relationnel, les graphes ou les séries temporelles. Grâce à cette flexibilité, les tenseurs peuvent jouer le rôle de modèle pivot pour traiter des ensembles de données hétérogènes [6]. Cela facilite leur utilisation dans de nombreux cas d'usage.

Les tenseurs sont munis d'opérateurs de décompositions tensorielles [9], comme Tucker ou CANDECOMP/PARAFAC [5], qui permettent de conduire des analyses exploratoires sur des données, notamment pour détecter des communautés ou des singularités [3]. Ces décompositions ont l'avantage de nécessiter peu de paramètres (en général, uniquement un rang est nécessaire). Le résultat d'une décomposition peut s'utiliser de multiple façons [4], de manière supervisée ou non.

Parmi ces décompositions, les décompositions couplées s'exécutent sur plusieurs tenseurs ayant au moins une dimension en commun [7]. Les tenseurs couplés présentent un fort intérêt puisqu'ils permettent de représenter conjointement des données à différents niveaux d'abstraction ou

à différents points de vue — par exemple, des données concernant les profils des utilisateurs dans un tenseur, et des données concernant les publications faites par ces utilisateurs dans un autre tenseur, ou encore en ayant un tenseur représentant des connaissances vis à vis de certains hashtags des publications. Dans cette configuration, le résultat des décompositions peut donc être affiné en considérant simultanément ces différents aspects.

Cette thèse consiste à explorer les décompositions tensorielles couplées, à la fois selon leurs capacités de modélisation et leurs capacités d'analyse. Pour ce faire, en plus des jeux de données publics, plusieurs jeux de données collectés sur Twitter lors de projets précédents sont à disposition, comme par exemple les discussions autour des élections présidentielles françaises de 2022 ou le COVID, contenant de plusieurs millions à plusieurs milliards de tweets. L'environnement matériel des serveurs de stockage et de traitement est opérationnel au Data Center Régional de l'UBFC, constituant un environnement adapté à l'expérimentation des propositions sur des données massives. Les travaux de thèse couvriront l'étude des techniques d'analyse de détection de robots dans les réseaux sociaux, la proposition d'une méthode basée sur les décompositions tensorielles couplées permettant de détecter les phénomènes de diffusion, le développement d'un prototype de cette proposition accompagné de sa validation expérimentale, ainsi que l'aide à l'interprétation des résultats en collaboration avec les chercheurs en sciences du langage et de la communication participant au projet de recherche.

Références

[1] Stefano Cresci. A decade of social bot detection. *Communications of the ACM*, 63(10) :72–83, 2020.

[2] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots : Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972, 2017.

[3] Annabelle Gillet, Éric Leclercq, and Nadine Cullot. Multi-level optimization of the canonical polyadic tensor decomposition at large-scale : Application to the stratification of social networks through deflation. *Information Systems*, 112 :102142, 2023.

[4] Annabelle Gillet, Éric Leclercq, and Lucile Sautot. A Guide to the Tucker Tensor Decomposition for Data Mining : Exploratory Analysis, Clustering and Classification. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems LIV : Special Issue on Data Management-Principles, Technologies, and Applications (TLDKS)*, pages 56–88. Springer, 2023.

[5] Annabelle Gillet, Éric Leclercq, and Lucile Sautot. The Tucker tensor decomposition for data analysis : capabilities and advantages. In *38ème Conférence sur la Gestion de Données (BDA)*, 2022.

[6] Annabelle Gillet, Éric Leclercq, Marinette Savonnet, and Nadine Cullot. Empowering big data analytics with polystore and strongly typed functional queries. In *Proceedings of the 24th Symposium on International Database Engineering & Applications (IDEAS)*, pages 1–10, 2020.

[7] Aditya Gudibanda, Tom Henretty, Muthu Baskaran, James Ezick, and Richard Lethin. All-at-once decomposition of coupled billion-scale tensors in Apache Spark. In *High Performance extreme Computing Conference*, pages 1–8. IEEE, 2018.

[8] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3) :455–500, 2009.

[9] Evangelos E Papalexakis, Christos Faloutsos, and Nicholas D Sidiropoulos. Tensors for data mining and data fusion : Models, applications, and scalable algorithms. *Transactions on Intelligent Systems and Technology (TIST)*, 8(2) :16, 2016.

Profil recherché

- étudiant en Master 2 avec spécialité informatique ;
- curiosité et rigueur scientifique ;
- compétences en gestion de données, algorithmique et programmation, des connaissance en Scala/Spark seraient appréciées ;
- bonne maîtrise de l'anglais écrit et parlé.

Financement ANR-23-CE38-0002-01

Dossier à envoyer avant le 30 juin 2024
Début du contrat : 1er octobre 2024
Durée du contrat : 36 mois
Salaire brut mensuel : 1944€

Encadrement de la thèse

Annabelle Gillet, Maître de Conférences, annabelle.gillet@u-bourgogne.fr – co-encadrante
Éric Leclercq, Professeur, eric.leclercq@u-bourgogne.fr – directeur

Pièces à fournir

Le dossier de candidature est à envoyer par mail à annabelle.gillet@u-bourgogne.fr et eric.leclercq@u-bourgogne.fr :

- CV ;
- lettre de motivation exposant votre intérêt et vos compétences pour le sujet proposé ;
- relevés de notes de la troisième année de Licence et du Master ;
- un mémoire réalisé pendant le cursus de Master démontrant vos capacités de rédaction ;
- lettre(s) de recommandation.

Thesis title: Coupled tensor decompositions for detecting diffusion phenomena in social networks

Host Laboratory

Laboratoire d'Informatique de Bourgogne (LIB EA 7534), University of Burgundy, Dijon

Specialty

Computer Science

Keywords

Data analysis, Tensor decompositions, Detection of diffusion phenomena

Job description

The goal of data analysis is to extract value from data. However, it is a complex process, that can use data from different sources, with different characteristics, stored in different systems relying on different data models, etc. Furthermore, the kind of analysis can enforce constraints restricting usable algorithms. This PhD thesis is part of the Beelzebot project (ANR-23-CE38-0002-01), aiming at detecting bot armies on Twitter, and at providing domain experts with interpretable results to produce alerts during disinformation campaigns. In this context, it is difficult to use supervised algorithms, because bots' behaviors rapidly change to avoid detection methods. It implies to regularly produce new train datasets, whereas domain experts have more difficulty distinguishing bots from human accounts [2]. Interactions and users' activity on social networks are becoming more and more used to detect bots' activity as diffusion phenomena rather than using only information about individual accounts [1].

Tensors are multi-dimensional mathematical objects [8], able to integrate data having different data models, such as the relational model, graphs or time series. Thanks to their flexibility, tensors can act as pivot model to process heterogeneous data [6], making them useful in various use cases.

Tensors are equipped with tensor decomposition operators [9], such as Tucker or CANDECOMP/PARAFAC [5], allowing to conduct exploratory analysis on data, for example to detect communities or singularities [3]. Tensor decompositions do not require many parameters, they often only need a rank. The result can be used in different ways [4], in supervised or unsupervised methods.

Among tensor decompositions, coupled decompositions take as input several tensors having at least one shared dimension [7]. Coupled tensors are highly interesting because they allow to represent jointly data at different abstraction levels — for example, data of users' profile in a tensor, and publications made by these users in an other tensor, or even a tensor representing knowledge regarding hashtags used in publications. Coupled tensors allow to refine decomposition results by taking into consideration the abstraction levels.

This PhD thesis consists in exploring coupled tensor decompositions, regarding their modelling and analytics capabilities. To do so, on top of public datasets, several datasets of millions to billions tweets collected during previous projects (regarding 2022 French presidential elections or COVID for example) are available. Several high performance servers are hosted at the regional Data Center, and provide an experimental environment to evaluate propositions on Big Data. PhD researches will include the study of bot detection techniques in social networks, the proposition of a method based on coupled tensor decomposition to detect diffusion phenomena, the realisation

of a prototype and its experimental validation, as well as helping domain experts to interpret the results of the method in collaboration with the researchers in language and communication that participate in the research project.

References

[1] Stefano Cresci. A decade of social bot detection. *Communications of the ACM*, 63(10):72–83, 2020.

[2] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, pages 963–972, 2017.

[3] Annabelle Gillet, Éric Leclercq, and Nadine Cullot. Multi-level optimization of the canonical polyadic tensor decomposition at large-scale: Application to the stratification of social networks through deflation. *Information Systems*, 112:102142, 2023.

[4] Annabelle Gillet, Éric Leclercq, and Lucile Sautot. A Guide to the Tucker Tensor Decomposition for Data Mining: Exploratory Analysis, Clustering and Classification. In *Transactions on Large-Scale Data-and Knowledge-Centered Systems LIV: Special Issue on Data Management-Principles, Technologies, and Applications (TLDKS)*, pages 56–88. Springer, 2023.

[5] Annabelle Gillet, Éric Leclercq, and Lucile Sautot. The Tucker tensor decomposition for data analysis: capabilities and advantages. In *38ème Conférence sur la Gestion de Données (BDA)*, 2022.

[6] Annabelle Gillet, Éric Leclercq, Marinette Savonnet, and Nadine Cullot. Empowering big data analytics with polystore and strongly typed functional queries. In *Proceedings of the 24th Symposium on International Database Engineering & Applications (IDEAS)*, pages 1–10, 2020.

[7] Aditya Gudibanda, Tom Henretty, Muthu Baskaran, James Ezick, and Richard Lethin. All-at-once decomposition of coupled billion-scale tensors in Apache Spark. In *High Performance extreme Computing Conference*, pages 1–8. IEEE, 2018.

[8] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[9] Evangelos E Papalexakis, Christos Faloutsos, and Nicholas D Sidiropoulos. Tensors for data mining and data fusion: Models, applications, and scalable algorithms. *Transactions on Intelligent Systems and Technology (TIST)*, 8(2):16, 2016.

Applicant profile

- student with Master degree (or equivalent) in Computer Science;
- curiosity and scientific rigor;
- data management, algorithmic and programming, knowledge with Scala/Spark would be greatly appreciated;

- good writing and reading skills in English;
- French level B1 mandatory (interdisciplinary project involving researchers in social, communication and language sciences).

Funding ANR-23-CE38-0002-01

Documents to be sent by June 30, 2024
Start of contract: October 1, 2024
Contract duration: 36 months
Gross monthly salary: 1944€

Thesis advisors

Annabelle Gillet, Assistant professor, annabelle.gillet@u-bourgogne.fr – co-supervisor
Éric Leclercq, Professor, eric.leclercq@u-bourgogne.fr – supervisor

Application documents

Documents must be sent to annabelle.gillet@u-bourgogne.fr and eric.leclercq@u-bourgogne.fr:

- CV;
- cover letter demonstrating your interest and abilities for the proposed thesis;
- transcripts and results of Bachelor, Master or equivalent degrees;
- a memoir/dissertation/Master thesis written during the last two years of studies to show redaction capabilities;
- reference letter(s).