Défi EGC 2020 : Analyse tensorielle de données issues de la conférence EGC

Rafika Boutalbi [*,**], Lazhar Labiod [*], Mohamed Nadif[*]

[*]Lipade, Université de Paris, 75006 Paris, France
[**]Trinov, Paris, France
<prénom.nom>@parisdescartes.fr

Résumé. La conférence EGC attire chaque année un nombre important de chercheurs dans le domaine de l'extraction et la gestion des connaissances. Cette année est organisée la 20 ème édition de la conférence EGC et la seconde édition du défi EGC qui a pour challenge d'analyser la dynamique de l'évolution de la conférence. Ce travail présente une analyse originale basée sur une approche tensorielle intégrant plusieurs sources de données dans un objectif d'analyse de thématiques, des communautés d'auteurs et des recommandations.

1 Introduction

La conférence EGC compte aujourd'hui parmi les conférences françaises qui attirent le plus grand nombre de chercheurs chaque année. Pour la 20ème édition, la conférence relance la deuxième édition du Défi EGC afin d'analyser et prédire l'évolution de la conférence depuis 2001. Pour cela différentes sources de données ont été mises à disposition des participants.

Nous proposons dans ce travail une approche multi-dimensionnelle permettant de considérer différentes sources de données (voir figure 1) et d'extraire des informations intéressantes. Pour ce faire nous avons réalisé dans un premier temps un prétraitement des données ensuite nous avons structuré les données en tenseurs afin de combiner différentes informations. Ce travail se compose de trois grandes parties (i) L'extraction des thématiques contenues dans les articles publiés dans la conférence EGC avec l'analyse de l'aspect temporel (ii) l'analyse des communautés d'auteurs (iii) la recommandation des évaluateurs pour le comité de programme (iv) Enfin l'étude de l'évolution de la popularité de la conférence en considérant le réseau social Twitter.

L'article suivant s'organise comme suit : La section 2 présente les différents pré-traitements réalisés ainsi que les variables retenues pour chaque source de données. La section 3 présente brièvement le modèle proposé pour le clustering de tenseurs. La section 4 est consacrée à l'analyse de thématiques. Dans la section 5, les communautés d'auteurs sont analysées. Dans la section 6 nous décrivons le système de recommandation des évaluateurs ainsi que les résultats obtenus. La section 7 est dédiée à l'étude de l'évolution de la popularité de la conférence dans le temps. Enfin la section 8 conclue notre contribution.