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.

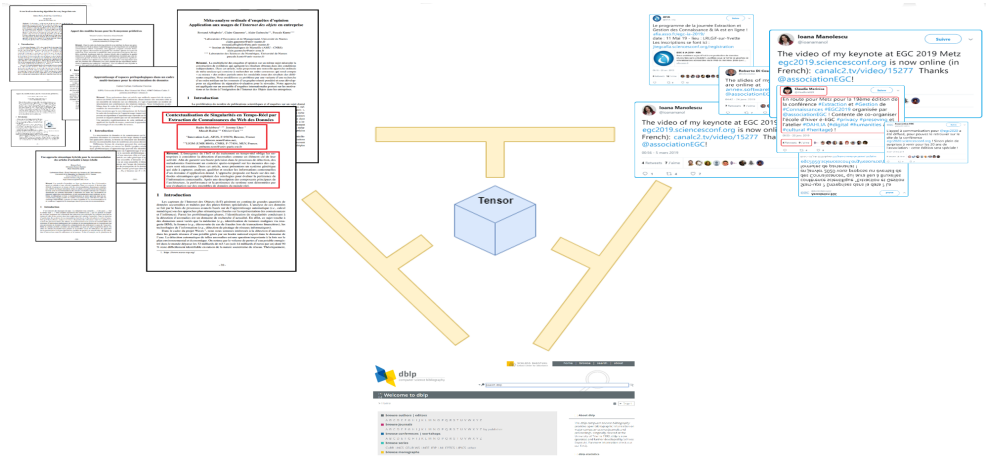

FIG. 1: Objectif du travail proposé

2 Traitement et description des données

Afin d'analyser l'historique de la conférence EGC, trois sources de données sont à notre disposition (i) La liste de tous les articles ainsi que leur contenu, (ii) Les données issues du réseau social Twitter, et (iii) Les emails des listes de diffusion EGC. Nous avons réalisé un ensemble de pré-traitements et sélectionné les variables les plus pertinentes pour notre étude. Un ensemble de données exogènes a également été introduit afin de nous aider à l'interprétation des résultats.

2.1 Données articles

Après la phase de pré-traitement, nous avons extrait un ensemble d'informations intéressantes :

- Les titres des articles.
- Les résumés des articles.
- La liste des auteurs pour chaque article.
- La liste des références citées par chaque article, et qui ont été extraites à partir du document PDF de l'article.

En considérant seulement les articles pour lesquels nous disposons de toutes ces informations (soit 1096 articles) plusieurs matrices de données ont été construites :

- La matrice des Documents-termes en considérant le titre : Cette matrice a été construite à partir des titres des documents $\mathcal{T}$ où chaque cellule $\mathcal{T}[i, j]$ représente l'occurrence du mot j dans le document i.
- La matrice des Documents-termes en considérant le résumé. Cette matrice $\mathcal{R}$ a été construite de la même manière que la matrice $\mathcal{T}$.
- La matrice Documents-auteurs $\mathcal{A}$ où chaque cellule $\mathcal{A}[i, j]$ est égale à 1 si l'auteur a écrit l'article, 0 sinon.
- La matrice Documents-références $\mathcal{F}$ où chaque cellule est égale à 1 si la référence a été citée dans l'article, 0 dans le cas contraire.

— La matrice Auteurs-affiliation $\mathcal{H}$ où chaque cellule $\mathcal{H}[i,j]$ est égale à 1 si l'auteur i appartient à l'institution j et 0 dans le cas contraire.
— La matrice Auteurs-termes $\mathcal{B}$ a été construite à partir de la matrice $\mathcal{T}$ binarisée et la matrice $\mathcal{A}$ à l'aide de $\mathcal{A}^T\mathcal{T}$ où chaque cellule $\mathcal{B}[i,j]$ représente le nombre de fois que le terme j a été utilisé par l'auteur i.

2.2 Données Twitter

À partir de l'API Twitter nous avons extrait tous les tweets liés à la conférence EGC à travers les mots-clés. Le nombre de tweets, le nombre de partages et le nombre de likes. Nous avons ensuite extrait ces mêmes données mais pour d'autres conférences internationales notamment ACL.

2.3 Données exogènes récupérées du Web

Nous avons d'abord récupéré à partir de l'API DBLP toutes les informations concernant les publications précédentes pour tous les auteurs notamment les titres des publications. Pour enrichir notre analyse des communautés d'auteurs, la variable *sexe* des auteurs a été également récupérée.

3 Modèle proposé : Co-clustering de tenseurs

Le traitement des données relationnelles repose généralement sur des outils de modélisation des données relationnelles. Un graphe non orienté peut représenter ces données avec des sommets représentant des entités et des arêtes décrivant les relations entre les entités. Ces relations peuvent être bien décrites par plusieurs graphes non orientés sur le même ensemble de sommets avec des arêtes issues de graphes différents capturant des relations hétérogènes. Les sommets de ces réseaux sont souvent structurés en classes inconnues avec différentes propriétés de connectivité. Pour extraire des classes pertinentes, nous proposons une méthode de co-clustering appropriée dérivant d'un modèle probabiliste prenant en compte plusieurs graphes se présentant sous forme d'un tenseur.

3.1 Modèle Poissonien des blocs latents (PLBM)

Dans le modèle PLBM (Govaert et Nadif, 2018) nous considérons une fonction de densité de probabilité conditionelle de $\mathbf{X}$ définie par $\prod_{i,j,k,\ell}\{\mathcal{P}(x_{ij}; x_{i.}x_{.j}\gamma_{k\ell})\}^{z_{ik}w_{j\ell}}$ où $x_{i.} = \sum_j x_{ij}$ et $x_{.j} = \sum_i x_{ij}$. Avec cette hypothèse, nous considérons alors le modèle LBM, où les deux ensembles I et J sont considérés comme des échantillons aléatoires, et les étiquettes des lignes et des colonnes deviennent des variables latentes. Par conséquent, le paramètre du modèle de bloc latent est $\mathbf{\Omega} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\gamma})$, avec $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ et $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_m)$ où $(\pi_k = P(z_{ik} = 1), k = 1, \ldots, g)$, $(\rho_\ell = P(w_{j\ell} = 1), \ell = 1, \ldots, m)$ sont les proportions du mélange et $\boldsymbol{\gamma} = (\gamma_{k\ell}; k = 1, \ldots g, \ell = 1, \ldots, m)$. En supposant que les données complétées sont le vecteur $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$, c'est-à-dire que nous supposons que les variables latentes $\mathbf{Z}$ et $\mathbf{W}$ sont connues, ainsi la log-vraisemblance complétée des données du modèle LBM peut être

écrite comme suit :

$$L_C(\mathbf{Z}, \mathbf{W}, \mathbf{\Omega}) = \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_\ell + \sum_{i,j,k,\ell} z_{ik} w_{j\ell} \log \mathcal{P}(x_{ij}; x_{i.} x_{.j} \gamma_{k\ell}).$$

En notant $\mathcal{Z}$ et $\mathcal{W}$ les ensembles des partitions possibles $\mathbf{Z}$ pour I et $\mathbf{W}$ pour J, la probabilité $L_C(\mathbf{\Omega})$ des données observées est obtenue par somme sur $\mathcal{Z}$ et $\mathcal{W}$ de $\mathbf{Z}$ et $\mathbf{W}$. Cependant, cette somme n'est pas traitable. Dans (Govaert et Nadif, 2005), les auteurs ont proposé une approche variationnelle pour l'estimation des paramètres ; elle s'appuie sur la maximisation de

$$L_C(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{W}}, \mathbf{\Omega}) + H(\widetilde{\mathbf{Z}}) + H(\widetilde{\mathbf{W}}) \tag{1}$$

où $L_C(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{W}}, \mathbf{\Omega})$ est la vraisemblance floue des données complétées. $H(\widetilde{\mathbf{Z}}) = -\sum_{i,k} \widetilde{z}_{ik} \log \widetilde{z}_{ik}$ avec $P(z_{ik} = 1|\mathbf{X}) = \widetilde{z}_{ik}$, et $H(\widetilde{\mathbf{W}}) = -\sum_{j,\ell} \widetilde{w}_{j\ell} \log \widetilde{w}_{j\ell}$ avec $P(w_{j\ell} = 1|\mathbf{X}) = \widetilde{w}_{j\ell}$.

3.2 Modèle parcimonieux Poissonien des blocs latents (SPLBM)

Malgré cette paramétrisation efficace, le modèle PLBM peut souffrir de la sparsité. Récemment, dans (Ailem et al., 2017b,a), les auteurs ont proposé une variante pour le co-clustering de matrices documents-termes sparse. Avec SPLBM, les auteurs supposent que pour chaque bloc diagonal kk les valeurs $x_{ij} \sim$ Poisson (λ_{ij}) où $\lambda_{ij} = x_{i.} x_{.j} \sum_k [z_{ik} w_{jk}] \gamma_{kk}$. Deuxièmement, ils supposent que pour chaque bloc $k\ell$ avec $k \neq \ell$ $x_{ij} \sim$ Poisson (λ_i) où le paramètre λ_{ij} prend la forme suivant : $\lambda_{ij} = x_{i.} x_{.j} \sum_{k,\ell \neq k} [z_{ik} w_{j\ell}] \gamma$.

3.3 SPLBM pour données tensorielles

Contrairement à la méthode LBM classique qui considère la matrice de données $\mathbf{X} = [x_{ij}] \in \mathbb{R}^{n \times d}$, notre proposition Tensor SPLBM considère la matrice de données 3D avec $\mathbf{X} = [\mathbf{x}_{ij}] \in \mathbb{R}^{n \times n \times v}$ où n est le nombre de nœuds et v le nombre de graphes (slices). En supposant l'indépendance par graphe, la fonction de densité conditionnelle de Poisson est donnée par $\prod_{i,j=1}^{n} \prod_{k=1}^{g} \prod_{b=1}^{v} \{\mathcal{P}(x_{ij}; x_{i.}^b x_{.j}^b \gamma_{k\ell}^b)\}^{z_{ik} w_{j\ell}}$. Comme $\mathbf{X}$ est symétrique par couche b, nous devons optimiser $\frac{1}{2}\mathcal{L}_C(\widetilde{\mathbf{Z}}, \widetilde{\mathbf{W}}, \mathbf{\Omega}) + H(\widetilde{\mathbf{Z}})$ qui prend la forme suivante

$$\begin{aligned} &\sum_{i,k} \widetilde{z}_{ik} \log \pi_k + \frac{1}{2} \sum_{i,j,k} \widetilde{z}_{ik} \widetilde{w}_{jk} \left(\sum_{b=1}^{v} \log \mathcal{P}(x_{ij}^b; x_{i.}^b x_{.j}^b \gamma_{kk}^b) \right) \\ &+ \frac{1}{2} \sum_{i \neq j, k \neq \ell} \widetilde{z}_{ik} \widetilde{w}_{j\ell} \left(\sum_{b=1}^{v} \log \mathcal{P}(x_{ij}^b; x_{i.}^b x_{.j}^b \gamma^b) \right) + H(\widetilde{\mathbf{Z}}). \end{aligned} \tag{2}$$

3.4 Algorithme `TSPLBM`

Pour estimer les paramètres du modèle, nous nous appuyons sur l'algorithme Variational EM (Govaert et Nadif, 2003, 2005, 2013) nous l'étendons au cas de plusieurs graphes se présentant sous forme d'un tenseur (Boutalbi et al., 2019). Dans la suite, L'algorithme appelé `TSPLBM` résume les deux étapes **E** et **M**. Sachant que $\sum_k \widetilde{z}_{ik} = \sum_k \widetilde{w}_{jk} = 1$, l'étape **E**,

consiste à calculer, pour tout i, j, k les probabilités a posteriori $\widetilde{z}_{ik}$ et $\widetilde{w}_{jk}$ étant donné les paramètres estimés $\mathbf{\Omega}$. Compte tenu des probabilités a posteriori précédemment calculées $\widetilde{\mathbf{Z}}$, l'étape **M** consiste à mettre à jour, $\forall k$, les paramètres π_k, $\boldsymbol{\gamma}^b_{kk}$ et $\boldsymbol{\gamma}^b$. Les paramètres estimés sont définis comme suit. Premièrement, en tenant compte des contraintes $\sum_k \pi_k = 1$, il est facile de montrer que $\pi_k = \frac{\sum_i \widetilde{z}_{ik}}{n}$. Deuxièmement, il est facile d'obtenir pour tous b, k

$$\boldsymbol{\gamma}^b_{kk} = \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{z}_{jk}x^b_{ij}}{\sum_i \widetilde{z}_{ik}x^b_{i.}\sum_j \widetilde{z}_{jk}x^b_{.j}} = \frac{x^b_{kk}}{[x^b_{k.}]^2} \text{et,}$$

$$\boldsymbol{\gamma}^b = \frac{N_b - \sum_{i,j,k} \widetilde{z}_{ik}\widetilde{z}_{jk}x^b_{ij}}{N_b^2 - \sum_k \sum_i \widetilde{z}_{ik}x^b_{i.}\sum_j \widetilde{z}_{jk}x^b_{.j}} = \frac{N_b - \sum_k x^b_{kk}}{N_b^2 - \sum_k [x^b_{k.}]^2}.$$

L'algorithme `TSPLBM` alterne les deux étapes décrites précédemment **E-M**. À la convergence, une partition dure **Z** est déduite à partir des $\widetilde{z}_{ik}$ en utilisant le principe du maximum a posteriori.

Algorithme 1 : `TSPLBM`

Input : $\mathbf{X} = [\mathbf{x}_{ij}] \in \mathbb{R}^{n\times n\times v}$, g, m.

Initialisation : $\widetilde{\mathbf{W}}^{(0)}$ et $\mathbf{\Omega}^{(0)}$

repeat

  Étape E

    Mise à jour des $\tilde{z}_{ik}$

    $\widetilde{z}^{t+1}_{ik} \propto \pi_k + \frac{1}{2}\left(\sum_{j,k} \widetilde{z}^t_{jk} \sum_{b=1}^v \mathcal{P}^{ijb}_{kk} + \sum_{j\neq i, k\neq \ell} \widetilde{z}^t_{j\ell} \sum_{b=1}^v \mathcal{P}^{ijb}_{k\ell}\right)$,

    où $\mathcal{P}^{ijb}_{kk} = \log \mathcal{P}(x^b_{ij}; x^b_{i.}x^b_{.j}\gamma^b_{kk})$ **et** $k \neq \ell$ $\mathcal{P}^{ijb}_{k\ell} = \log \mathcal{P}(x^b_{ij}; x^b_{i.}x^b_{.j}\gamma^b)$.

  Étape M

    Mise à jour de $\mathbf{\Omega}$

until *convergence*;

return Z, $\mathbf{\Omega} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\gamma})$

4 Analyse de topics contenus dans les articles

Afin d'analyser les topics nous avons construit plusieurs graphes représentant différentes relations entre les documents ; les graphes construits sont les suivants :

1. La matrice des co-termes entre articles en considérant le titre : Cette matrice a été construite à partir de la matrice Documents-termes binarisée et qui est calculée par $\mathcal{T}\mathcal{T}^T$.
2. La matrice des co-termes entre articles en considérant le résumé : Est construite de la même manière que la matrice des co-termes titre, mais en considérant la matrice Documents-termes des résumés tel que $\mathcal{R}\mathcal{R}^T$.
3. la matrice co-auteurs : Cette matrice a été construite à partir de la matrice Documents-auteurs représentant les auteurs qui ont contribué à l'article en calculant $\mathcal{A}\mathcal{A}^T$.

4. La matrice co-références : Chaque cellule représente le nombre de références en commun entre deux articles. Cette matrice est obtenue par $\mathcal{F}\mathcal{F}^T$.

En considérant ces quatre relations nous avons appliqué l'algorithme `TSPLBM`. Nous avons utilisé un nombre de classes égal à 8 en se basant sur la mesure de la modularité. La figure 2 représente la réorganisation des nœuds des quatre graphes en utilisant le partitionnement obtenu par `TSPLBM`.

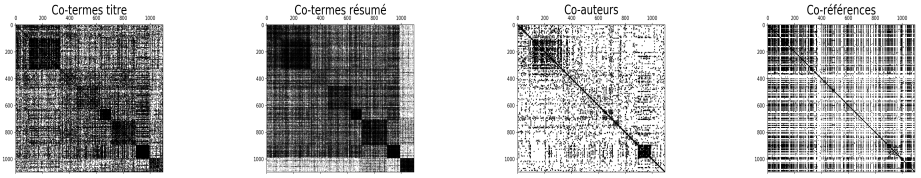

FIG. 2: Réorganisation des matrices d'adjacence des graphes de documents.

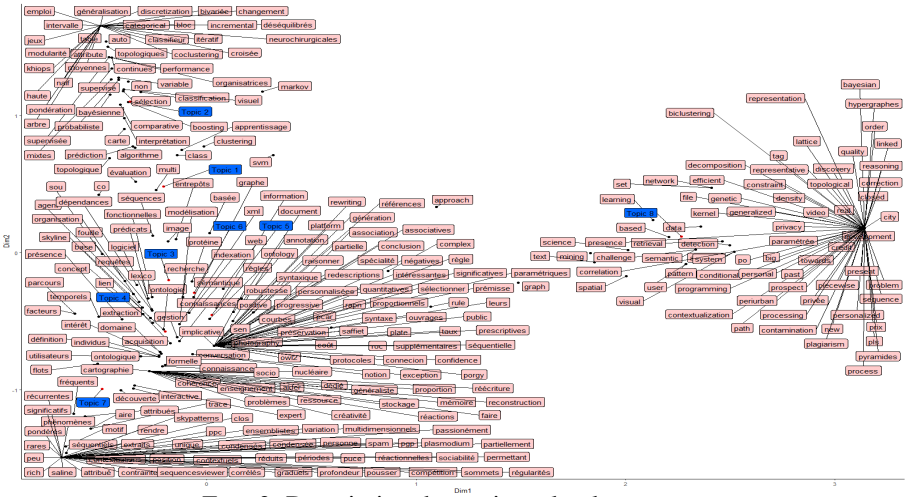

FIG. 3: Description des topics selon les termes.

Quelles sont les caractéristiques décrivant les topics des documents ? À partir des classes de documents obtenues ainsi que la matrices document-termes $\mathcal{T}$ et la matrice documents-auteurs $\mathcal{A}$, nous avons construit les matrices Topics-termes $\mathcal{D}$ et Topics-auteurs $\mathcal{G}$. La matrice topics-termes représente le nombre de fois ou les différents termes ont été cités dans ce topic. La matrice la Topics-auteurs rapporte le nombre d'articles qu'un auteur a publié dans ce topic.

Nous avons appliqué une analyse des correspondances (CA) (Benzecri, 1973) sur les matrices Topics-termes et Topics-auteurs. Le choix de la visualisation par la CA est justifié par le modèle Poissonien des blocs latents ; pour plus de détails voir (Govaert et Nadif, 2018). Les résultats obtenus sont illustrés dans la figure 3 (respectivement 5). La figure 4 présente les fréquences des termes par classe (topic) de documents. Nous pouvons décrire les 8 topics :

— Classe 1 ou topic 1 traite d'algorithmes de classification non supervisée.
— Classe 2 ou topic 2 traite des algorithmes d'apprentissage supervisé ou non supervisé avec toutes les problématiques liées à cette thématique.
— Classe 3 ou topic 3 traitant de l'extraction de connaissances.

— Classe 4 ou topic 4 traitant le thème des bases de connaissances et ontologies.
— Classe 5 ou topic 5 traite des règles d'association.
— Classe 6 ou topic 6 traite du web sémantique.
— Classe 7 ou topic 7 traite de l'extraction de patterns ou de motifs fréquents.
— Classe 8 ou topic 8 traite du datamining et regroupe en grande partie les articles en anglais. On remarque que ce topic se distingue le plus des autres sur la figure 3.

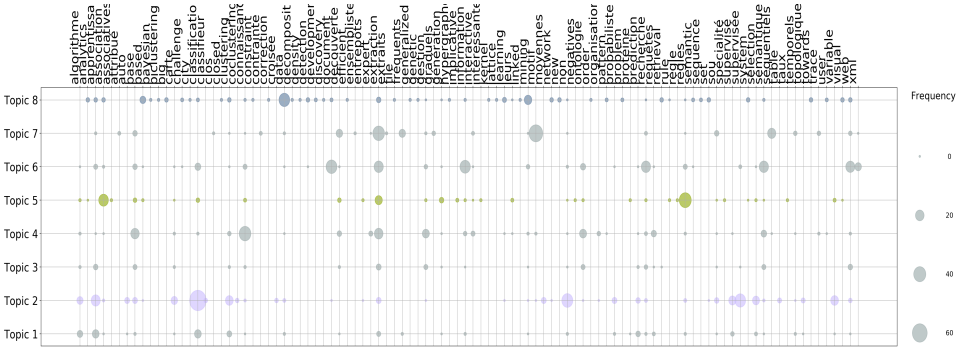

FIG. 4: Fréquences des termes qui ont le plus contribué à la CA par topics.

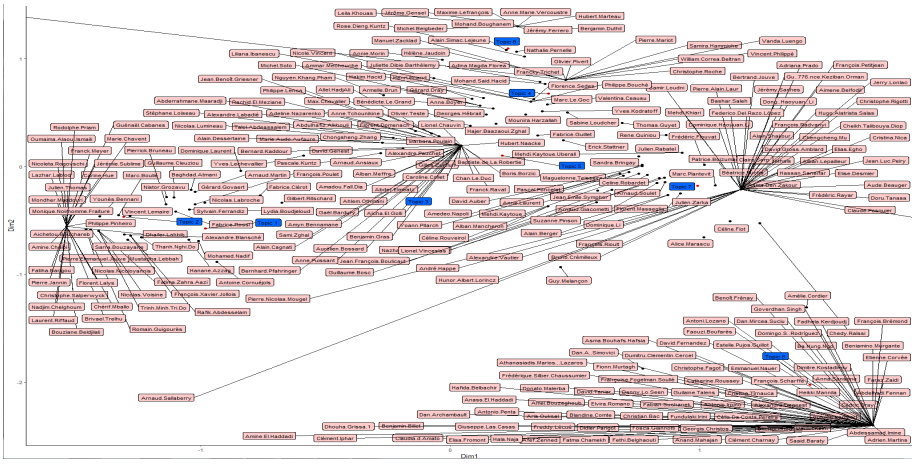

FIG. 5: Description des topics selon les auteurs.

Dans la figure 5 nous tentons de mettre en évidence l'intérêt de certains auteurs pour un topic donné. En effet, nous pouvons constater que chacun des topics est caractérisé par un ensemble de chercheurs qui y contribuent le plus. Pour topic 2 traitant de l'apprentissage automatique nous observons des auteurs comme *Vincent Lemaire*, *Marc Boullé*. Pour topic 7 abordant la détection des motifs fréquents, nous retenons par exemple *Marc Plantvit* et *Céline Robardet*. Enfin le topic 4 traitant des règles d'association est par contre illustré par des auteurs tels que *Florence Sedges*, *Marc Le Goc* et *Philippe Bouché*.

Comment évoluent ces thématiques dans le temps ? Nous proposons d'étudier l'évolution dans le temps des 8 topics retrouvés précédemment. Pour cela nous calculons le nombre d'ar-

ticles par topic et par année. La figure 6 présente l'évolution du nombre de publications par topic et par année. Nous avons affiché cette évolution pour trois topics différents notamment topic 2, topic 4 et topic 8. Nous pouvons remarquer que le topic 2 ne suit pas la tendance mais il est caractérisé par deux pics en 2008 et en 2014 alors que topic 4 a une tendance décroissante. Enfin topic 8 représente les articles anglophones avec une tendance croissante jusqu'à 2016 ce qui peut être expliqué par l'augmentation de la présence de chercheurs non francophones de 2001 à 2016.

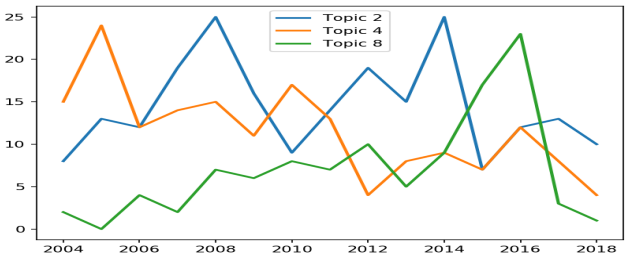

FIG. 6: Evolution des topics dans le temps.

5 Analyse des communautés d'auteurs

Afin d'analyser les communautés d'auteurs nous avons construit plusieurs graphes représentant différentes relations entre ces auteurs. Nous avons considéré les 816 auteurs pour lesquels nous avons pu récupérer l'affiliation. Les graphes construits sont les suivants :

1. Matrice des co-termes représentant le nombre de termes en commun qu'ont utilisé deux auteurs. Cette matrice a été construite à partir de la matrice Auteurs-termes binarisée et qui est calculée par $\mathcal{B}\mathcal{B}^T$.
2. Matrice co-auteurs : cette matrice a été construite à partir de la matrice Documents-auteurs représentant le nombre de documents qu'ont rédigé en commun deux auteurs en calculant $\mathcal{A}^T\mathcal{A}$.
3. Matrice co-affiliations : chaque cellule égale à 1 indique que deux auteurs appartiennent à une même institution et zéro dans le cas contraire. Elle est calculée par $\mathcal{H}\mathcal{H}^T$.
4. Matrice co-topics : cette matrice est construite à partir de la matrice topics-auteurs et est calculée par $\mathcal{G}\mathcal{G}^T$.

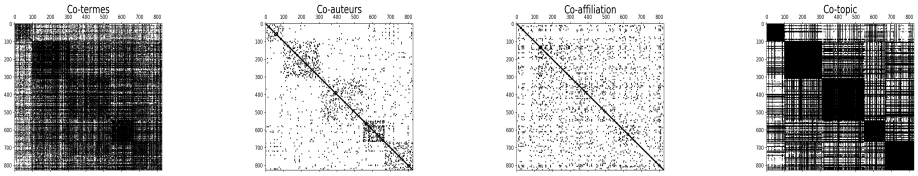

FIG. 7: Réorganisation des matrices d'adjacence des graphes d'auteurs.

En considérant ces quatre relations entre auteurs nous avons appliqué l'algorithme `TSPLBM`. Nous avons utilisé un nombre de classes égal à 5 en se basant sur la mesure de modularité. La

figure 7 représente la réorganisation des nœuds des quatre graphes en utilisant le partitionnement obtenu par TSPLBM.

Comment peut-on interpréter les communautés d'auteurs découvertes ? Le modèle proposé TSPLBM permet de combiner plusieurs informations et ainsi faciliter l'interprétation des résultats. Nous avons construit la matrice topics-affiliations et appliquer une CA sur cette dernière. La figure 8 visualise les résultats de la CA représentant les différentes communautés d'auteurs ainsi que les affiliations qui y contribuent le plus. Nous pouvons remarquer que communauté 1 représente en grande partie les chercheurs étrangers avec des noms de domaine tels que *@nac.ac.uk*, *@unicampania.it*, *@uni-konstonz.de*, etc.

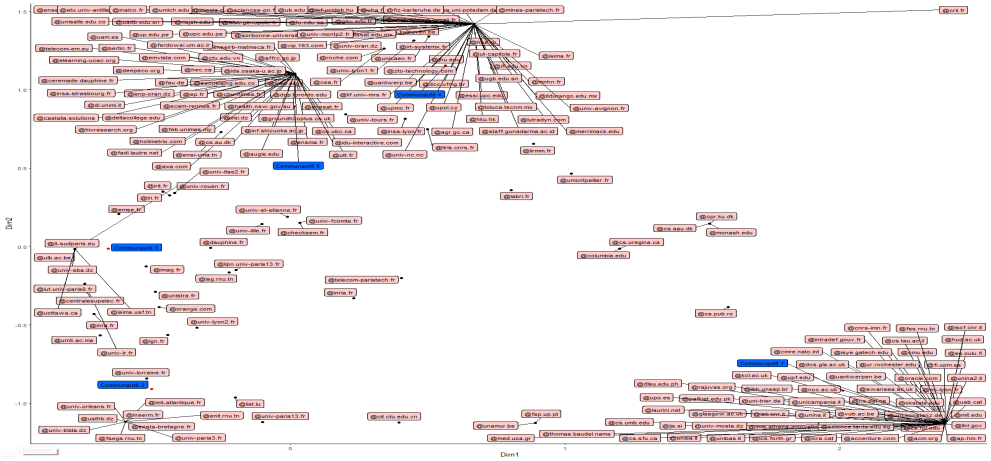

FIG. 8: Description des communautés d'auteurs par affiliation.

Y'a-t-il des communautés d'auteurs qui se distinguent par rapport à la parité Homme-Femme ? Nous avons calculé la proportion Homme-Femme par communauté d'auteurs. L'objectif est d'apprécier le *sex-ratio* et le niveau de parité homme-femme au sein des différentes communautés. Dans la figure 9 on observe la proportion Homme-Femme pour les cinq communautés d'auteurs. Il apparait que toutes les communautés présentent quasiment la même proportion d'hommes 78-80% et de femmes 20-22%.

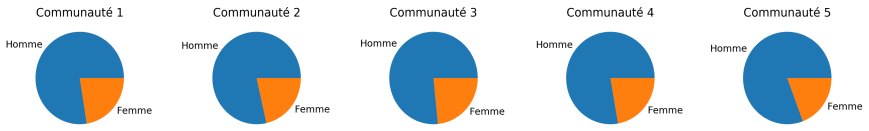

FIG. 9: Proportion Homme-Femme par communauté d'auteurs.

6 Recommandation des articles au comité de programme

Nous proposons dans cette partie un système de recommandation permettant de faciliter l'évaluation des articles soumis à la conférence EGC. Cette recommandation permet de pro-

poser un ensemble de chercheurs approprié pour l'évaluation d'un nouvel article. Il est donc plus simple de constituer un comité de programme en se basant sur ces recommandations. Les articles pour lesquels nous réalisons une recommandation d'évaluateurs sont les articles qui n'ont pas été considérés dans la première partie de détection de topics (section 4). La figure 10 représente le schéma de fonctionnement du système de recommandation proposé. Pour un nouvel article soumis à la conférence EGC, les étapes suivies sont :

— Produire la représentation vectorielle du titre de l'article.
— A partir de la matrice document-termes (titre) et des topics qui ont été extraits, soustraire la représentation vectorielle pour chaque topic.
— Calculer la similarité cosinus entre la représentation vectorielle du titre du nouvel article et la représentation vectorielle des topics. Ainsi nous pouvons attribuer le nouvel article à un des 8 topics.
— Une fois l'article attribué à un topic, nous sélectionnons les 30 auteurs qui publient le plus fréquemment dans ce topic, en considérant les données EGC.
— Afin d'améliorer la diversité et la pertinence des recommandations, nous récupérons, à l'aide de l'API DBLP, tous les travaux publiés disponibles des 30 auteurs sélectionnés.
— Nous pouvons construire une représentation vectorielle pour chacun des auteurs en se basant sur les titres de tous ses articles publiés et disponibles sur DBLP.
— Un score de similarité est calculé entre le titre du nouvel article et la représentation vectorielle des auteurs.
— Les trois auteurs ayant obtenu les plus grands scores de similarités sont recommandés pour évaluer l'article, à condition qu'ils n'appartiennent pas à la même institution de l'un des auteurs de l'article en question.

TAB. 1: Résultats des recommandations.

Titres	**Auteurs recommandés**
Analyse Ontologique de scénario dans un contexte Big Data	Fatiha Saïs , Stavrakas Yannis, Thomas Tamisier 0.293, 0.287, 0.284
Big Data for understanding human dynamics the power of networks	Stavrakas Yannis, Thomas Tamisier, Raja Chiky 0.331, 0.328, 0.317
Community structure in complex networks	Faraz Zaidi , Christine Largeron, Guy Melançon 0.284, 0.19, 0.141
Détection de Singularités en temps-réel par combinaison d'apprentissage automatique et web sémantique basés sur Spark	Alain Simac-Lejeune, Jérémy Ferrero, Thierry Despeyroux 0.269, 0.194, 0.174
eDOI : exploration itérative de grands graphes multi-couches basée sur une mesure de l'intérêt de l'utilisateur	David Genest, Djamel Abdelkader Zighed, Jean-Benoît Griesner 0.125, 0.119, 0.093
Fouille de Motifs Graduels Fermés Fréquents Sous Contrainte de la Temporalité	Lionel Vinceslas, Jean-Emile Symphor, François Rioult 0.156, 0.124 , 0.105
Long-range influences in (social) networks	Nacéra Bennacer, Christine Largeron , Rim Faiz 0.189, 0.148, 0.136
Méthode d'Apprentissage pour Extraire les Localisations dans les MicroBlogs	Emmanuel Viennet, Isabelle Tellier, Marc Boullé 0.174, 0.157, 0.093

Nous présentons dans Table 1 quelques exemples de résultats de recommandations. Par exemple, si on considère l'article « Analyse Ontologique de scénario dans un contexte Big Data», le système de recommandation propose trois évaluateurs qui sont *Fatiha Saïs*, *Stavrakas Yannis* et *Thomas Tamisier* avec des score de 0.293, 0.287 et 0.284 respectivement. Nous remarquons en effet que ces chercheurs travaillent bien dans la thématique des ontologies et du web sémantique, deux domaines connexes au contenu de l'article en question.

7 Etude de l'évolution de la popularité de la conférence EGC

Nous souhaitons étudier s'il y a une corrélation entre le nombre de soumissions (qui représente la popularité d'une conférence) et l'activité de cette dernière sur les réseaux sociaux,

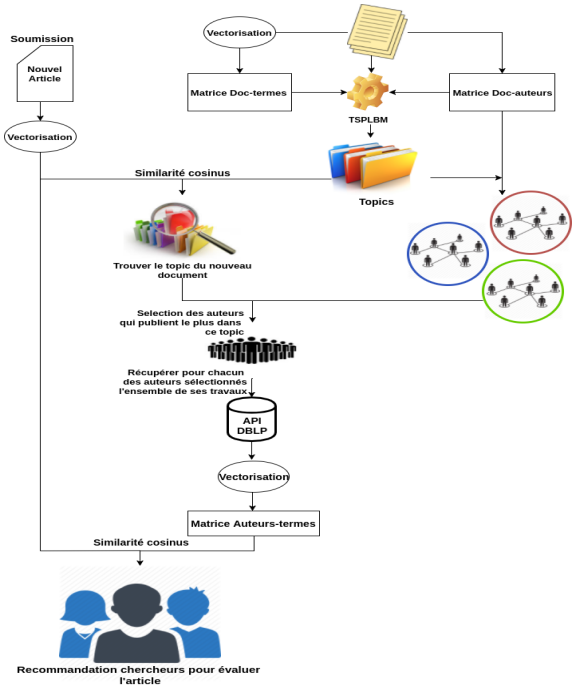

FIG. 10: Système de recommandation pour l'évaluation des articles.

en l'occurrence Twitter. Pour illustrer notre propos, nous avons choisi la conférence ACL (Association for Computational Linguistics). Nous avons récupéré les données de l'activité de la conférence EGC et ACL (à l'aide de l'API Twitter) ainsi que le nombre de soumissions annuelle pour la conférence ACL. La figure 11 présente les courbes de trois variables *nombre de tweets* EGC, *nombre de tweets* et *nombre de soumissions* à la conférence ACL. Nous remarquons que le nombre de soumission à l'année $(t+1)$ croit avec l'augmentation du nombre de tweets à l'année (t) ; ceci a été aussi constaté pour d'autres conférences prestigieuses. Cela suggère qu'une activité importante sur Twitter pourrait permettre d'avoir plus de visibilité de EGC et donc d'attirer plus de chercheurs.

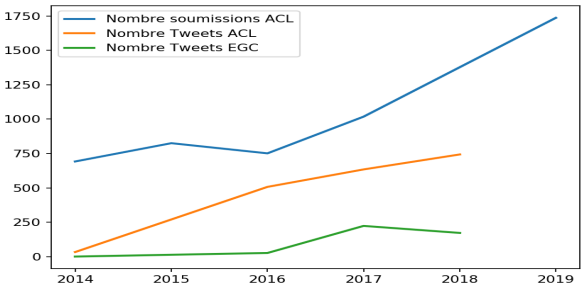

FIG. 11: Evolution du nombre de tweets et de soumissions.

8 Conclusion

Nous avons réalisé une étude basée principalement sur le modèle `TSPLBM` permettant d'analyser les données issues de EGC. Ainsi, l'analyse tensorielle des documents et des auteurs nous a permis d'extraire 8 topics et 5 communautés respectivement. Nous avons analysé et décrit les topics en utilisant les termes des titres, les résumés et les auteurs qui contribuent de manière importante. De la même manière les communautés d'auteurs ont été analysées en utilisant les affiliations et des variables exogènes telle que la variable *sexe* des auteurs.

Dans un deuxième temps nous avons présenté un système de recommandations d'évaluateurs afin de proposer une liste de chercheurs appropriés pour l'évaluation d'un article donné. Les résultats obtenus mettent en évidence la pertinence des recommandations obtenues. Enfin une analyse sur la corrélation entre le nombre de soumissions et l'activité d'une conférence sur les réseaux sociaux met en avant l'importance de la visibilité sur le réseau social Twitter.

Références

Ailem, M., F. Role, et M. Nadif (2017a). Model-based co-clustering for the effective handling of sparse data. *Pattern Recognition 72*, 108–122.

Ailem, M., F. Role, et M. Nadif (2017b). Sparse poisson latent block model for document clustering. *IEEE TKDE 29*(7), 1563–1576.

Benzecri, J.-P. (1973). *L'analyse des données, tome 2 : l'analyse des correspondances*. Paris : Dunod.

Boutalbi, R., L. Labiod, et M. Nadif (2019). Sparse tensor co-clustering as a tool for document categorization. In *SIGIR*, pp. 1157–1160.

Govaert, G. et M. Nadif (2003). Clustering with block mixture models. *Pattern Recognition 36*, 463–473.

Govaert, G. et M. Nadif (2005). An EM algorithm for the block mixture model. *IEEE Transactions on Pattern Analysis and machine intelligence 27*(4), 643–647.

Govaert, G. et M. Nadif (2013). *Co-clustering : models, algorithms and applications*. John Wiley & Sons.

Govaert, G. et M. Nadif (2018). Mutual information, phi-squared and model-based co-clustering for contingency tables. *Advances in Data Analysis and Classification 12*(3), 455–488.

Summary

The EGC conference attracts a significant number of researchers each year in the field of knowledge discovery and data mining (KDD). This year is organized the 20th edition of the conference and the 2nd edition of the EGC challenge about the conference evolution. The goal is to analyze the dynamics of the conference evolution. This work presents an original analysis based on a tensorial approach integrating several sources of data to investigate topics, authors' communities, and the attractiveness of EGC.