Défi EGC 2020 : Visualisation et embeddings de mots pour détecter les thématiques récentes

Philippe Suignard*

*EDF R&D, 7 boulevard Gaspard Monge, 91120 Palaiseau
philippe.suignard@edf.fr

Résumé. Dans cet article, nous montrons comment des techniques de type word et document embeddings, appris sur le corpus des articles de la conférence EGC, associés à une technique de visualisation de graphe, permettent de détecter les thématiques en émergence de cette conférence.

1 Introduction

Pour sa 20ème édition qui aura lieu en janvier 2020, EGC organise un défi, dont l'objectif « est de faire le bilan de l'évolution de la communauté EGC ces 20 dernières années et tenter d'en prédire l'avenir. Le principe est d'appliquer des techniques d'extraction et de gestion de connaissances afin d'expliquer la structure et l'évolution de l'ensemble des données au fil des années ». Il se trouve qu'un précédent défi (en 2016) avait été organisé sur le même sujet. Nous avons pris le parti de regarder ce qu'il y avait de nouveau depuis ce moment.

Les techniques d'embeddings de mots ont connu beaucoup de succès ces dernières années (Mikolov et al., 2013), mais sont généralement entraînées sur des corpus de données très volumineux. Nous voulons montrer ici, que cette technique peut être utilisée et entraînée sur des corpus de faible taille comme c'est le cas pour ce défi constitué d'un corpus d'un millier de documents environ.

2 Les données

Les données fournies par EGC sont constituées des titres des articles issus de la conférence EGC, de leur résumé, de leurs auteurs, d'un lien vers la 1ère page de l'article et d'un lien vers l'article (en PDF). Une partie des données est rédigée en anglais : titre et/ou résumé soit au final 1269 documents répartis sur 15 ans (de 2004 à 2018). Pour ne pas perturber les traitements, les parties en anglais sont éliminées, comme l'ont fait (Guille et al., 2016). Pour nos traitements, un document sera constitué par la concaténation du titre et du résumé (en ne gardant que ce qui est rédigé en français). Au final, nous disposons de **1179** documents.

Résumé. La détection d'influenceurs dans les réseaux sociaux s'appuie généralement sur une structure de graphe représentant les utilisateurs et leurs interactions. Récemment, cette tâche a tenu compte, en sus de la structure du graphe, du contenu textuel généré par les utilisateurs. Notre approche s'inscrit dans cette lignée : des informations sont extraites du contenu textuel par des règles linguistiques puis sont intégrées dans un système d'apprentissage automatique. Nous montrerons le prototype développé et son interface de visualisation qui facilite l'interprétation des résultats.

FIG. 1 – *Exemple de résumé.*

2.1 Correction orthographiques des données

Les données ont été converties directement par EGC du format PDF vers le format Texte. Mais elles contiennent des erreurs d'orthographes réparties principalement en deux catégories :

— Les mots collés ;
— Les mots contenant un tiret.

L'exemple présenté dans la figure 1 illustre les deux problèmes rencontrés dans les documents (ex : ID72, `http://editions-rnti.fr/render_pdf.php?p=1002325`). Le dernier mot d'une ligne est parfois concaténé avec le premier mot de la ligne suivante, ce qui génère des mots comme : « **généra-lement** » et « **cettelignée** » [1].

Les mots qui contiennent un tiret sont difficiles à corriger car il faut les différencier des « vrais » mots composés d'un tiret comme : « multi-labels », « cartes auto-organisatrices », « temps-réel » ou « spatio-temporel ». Dans ce cas, il est préfèrable de ne rien corriger pour éviter d'introduire du bruit. Pour les mots collés, nous réalisons les traitements suivants à l'aide du correcteur Hunspell (Németh et al., 2004) :

— Hunspell va détecter les mots mal orthographiés (selon les mots et les règles décrits dans des fichiers de paramétrage) ;
— Pour les mots mal orthographiés, Hunspell va proposer une liste de substituants possibles ;
— S'il propose des substituants composés de deux mots, nous remplaçons le mot initial par les deux mots [2] ;
— S'il ne propose pas de de substituants composés de deux mots, le mot initial est laissé tel quel.

Au final, cette méthode va apporter **4533 correction**s : « nousproposons » en « nous proposons » 18 fois, « extractionde » en « extraction de » 10 fois, etc.

1. Le résumé transformé au format texte par EGC est le suivant : « La détection d'influenceurs dans les réseaux sociaux s'appuie **généra-lement** sur une structure de graphe représentant les utilisateurs et leurs **interac-tions**. ... Notre approche s'inscrit dans **cettelignée** : des informations sont extraites du contenu textuel par des règles **linguistiques**, etc. »

2. Pour le mot « cettelignée », il proposera : « cette lignée », « cette-lignée » et « interlignée ». C'est la première proposition qui sera choisie. Parfois il peut y avoir plusieurs propositions. Pour le mot « descriptivesou », il proposera : « descriptive sou », « descriptive-sou », « descriptives ou », « descriptives-ou », « descriptive », « descriptivisme » et « prescriptive ». Dans ce cas, on remplacera « **descriptivesou** » par « **descriptive sou descriptives ou** ».

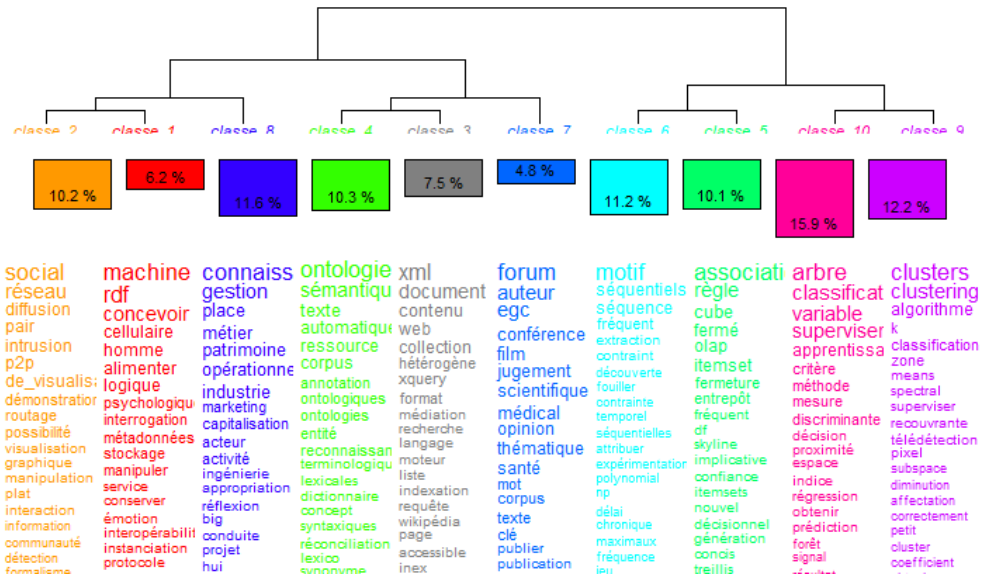

FIG. 2 – *Répartition des articles en thématique.*

3 Aperçu global des thématiques

Pour avoir un aperçu global des thématiques présentes dans les données, nous avons utilisé le logiciel Iramuteq (Ratinaud, 2009). Pour cela, les textes sont considérés dans leur ensemble sans les couper en segments. Le logiciel Iramuteq permet de réaliser des classifications non supervisées (clustering) grâce à une classification descendante hiérarchique décrite dans (Reinert, 1983) et (Reinert, 1986) en fractionnant le corpus de textes en deux parties de manière récursive, jusqu'à obtenir le nombre de classes désirées. Il ne s'appuie pas sur des techniques de type « embeddings » mais sur une approche de type « sac de mots ». La figure 2 présente les grandes catégories ou thématiques obtenues, très proches de celles constatées dans le défi EGC 2016 par (Guille et al., 2016) et (Cabanac et al., 2016), à savoir :

1. Réseau sociaux et visualisation
2. Interaction homme-machine : interaction, démonstration, interface, interagir, graphique, manipulation, etc.
3. Gestion des connaissances
4. Ontologie, sémantique, corpus, données textuelles
5. XML, document, web
6. Une catégorie orientée métier ou use-case : santé, détection d'opinions, film, forums, etc.
7. Motifs/séquences fréquent(e)s, fouille, extraction, découverte, etc.
8. Règles d'associations, entrepôt de données, etc.

9. Arbres de décision, classification supervisée, apprentissage, régression, prédiction, etc.
10. Cluster, clustering, algorithme, k-means, etc.

4 Analyse des quatre dernières années

Comme la période 2004-2015 a été couverte par le précédent défi, nous nous intéressons ici à la période la plus récente afin de découvrir les évolutions des thématiques. Pour cela, le corpus est séparé en 2 parties : 2004-2014 (77% du corpus total) et 2015-2018 (23% du corpus total). L'objectif consiste à utiliser des techniques d'embeddings de mots/documents et des techniques de visualisation pour détecter les nouveautés.

4.1 Prétraitements

Un modèle d'embeddings de mots est entraîné avec Word2Vec sur les articles (titre + résumé) avec les paramètres suivants : une fenêtre de 2 mots à gauche et 2 mots à droite, une couche cachée de taille 200, une architecture skip-gram, une fréquence minimale de 5 et 1000 itérations.

Les embeddings de documents sont ensuite calculés à partir des embeddings de mots. Les méthodes utilisées sont **Swem-aver** pour « Simple Word Embedding Model – average » (Shen et al., 2018) qui effectue une simple moyenne des vecteurs mots du document et **DoCoV** pour « Document Co Variance » (Torki, 2018) qui calcule une représentation vectorielle du document à partir de la matrice de covariance des embeddings de mots. Ces méthodes ont montré qu'elles pouvaient fournir des représentations vectorielles de qualité (Suignard et al., 2019). La suite du document présente les résultats obtenus avec la méthode DoCoV.

A partir des embeddings de documents, une similarité deux à deux est calculée entre tous les documents du corpus, grâce au cosinus.

4.2 Visualisation

Un fichier au format Gephi (Bastian et al., 2009) est créé en transformant chaque document en un point, un lien étant apposé entre deux documents si la similarité entre ces deux documents est supérieure à un seuil fixé. Ce seuil est fixé assez bas dans un premier temps pour que le réseau soit très interconnecté et le calcul du PageRank possible. La taille des points-documents est spécifiée par un calcul de PageRank (Page et al., 1999). Deux couleurs possibles sont attribuées aux points (rouge pour la période 2004-2014 et verte pour la période 2015-2018).

Une fois le fichier chargé dans Gephi, l'algorithme « Force-atlas 2 » est utilisé pour positionner les points dans le plan. Comme le seuil de similarité est fixé très bas, les points constituent une sorte de magma. L'utilisateur va ensuite augmenter interactivement et itérativement le seuil de similarité en dessous duquel les liens entre deux points sont supprimés. Des points vont alors se désolidariser du « magma » central et s'en éloigner pour se regrouper avec d'autres points sous la forme d'amas. Cette opération est répétée jusqu'à ce que suffisamment d'amas de points émergent, comme le montre la figure 3 (avec un seuil de similarité fixé à 0,3).

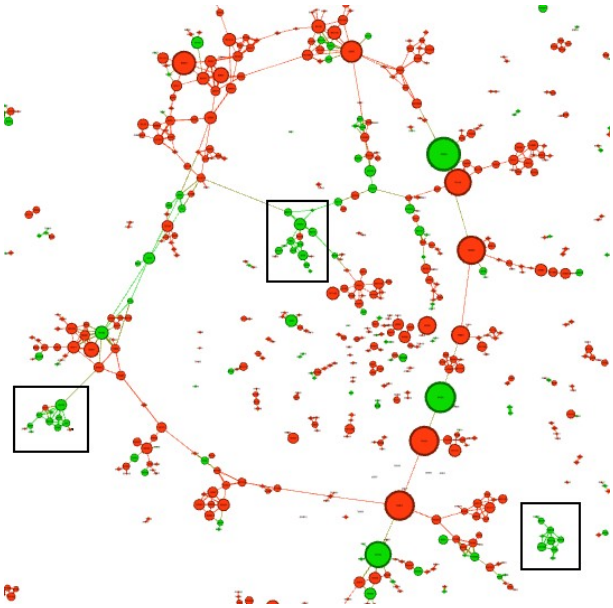

FIG. 3 – *Visualisation des amas de documents, en rouge pour la période 2004-2014 et en vert pour la période 2015-2018. Les zones encadrées sont décrites au paragraphe suivant.*

4.3 Résultats

L'analyse des regroupements de points ou amas permet de distinguer plusieurs cas de figures :

— Des regroupements de points tous rouges (ou très majoritairement rouge), qui signalent des documents ayant une thématique similaire entre eux, mais exclusivement présente dans la période 2004-2014, c'est-à-dire que la thématique a tendance à disparaître ;
— Des regroupements de points verts (ou très majoritairement verts), qui signalent une thématique présente uniquement à partir de 2015 et donc nouvelle ;
— Des groupes mixtes et équilibrés, preuve que la thématique perdure dans le temps.

Dans la suite de l'article, nous nous intéressons uniquement aux zones en émergence, c'est à dire aux 3 zones vertes présentées en détail sur la figure 4. Ces 3 groupes sont étiquetés manuellement en constatant qu'ils parlent d'une même thématique :

Premier groupe : le défi EGC 2016. Le premier groupe saillant (à gauche sur la figure 4) est constitué des 9 articles qui traitent du « Défi EGC 2016 ». Noter que l'ID148 ne mentionne ni « défi » ni « 2016 » dans son titre et dans sa synthèse, mais qu'il en fait bien partie. Idem pour l'ID169 qui traite de l'évolution des communautés au sein d'EGC. La détection de ce groupe valide globalement l'approche proposée puisque ce groupe est assez homogène (présence des mots « défi », « egc », « 2016 »), ce qui fait penser que la tâche est relativement facile, mais qu'il « ressemble » quand même au reste du corpus (dans la mesure où les documents du Défi parlent tous des thématiques EGC).

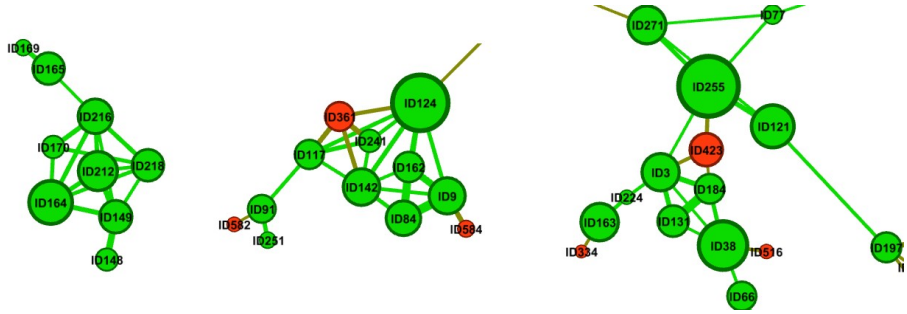

FIG. 4 – *Zoom sur trois zones nouvelles*

Deuxième groupe : classification multi-label. Le deuxième groupe saillant (au milieu sur la figure 4) concerne la classification **multi-label** :

- **ID124** : Sélection et transformation de variables pour la classification **Multi-Label** par une approche MDL : La classification **multi-label** est une extension de la classification supervisée au cas de plusieurs **labels**... ;
- **ID162** : Découverte de **labels dupliqués**... L'analyse des données comportementales ... traces... associées aux individus, ou **labels**. ... problème de dé duplication de **labels**... ;
- **ID9** : En classification **multi-labels**, chaque instance est associée à un ou plusieurs **labels**... ;
- **ID84** : La classification **multi-labels** graduée est la tâche d'affecter à chaque donnée l'ensemble des **labels**... ;
- **ID142** : Nous présentons un outil interactif de classification **multi label** développé au sein du groupe Orange... ;
- **ID117** : Nous décrivons dans cet article notre réponse au défi EGC 2017... Cela nous a permis de sélectionner les meilleurs classifieurs **uni-label** et **multi-label**. Autant sur la tâche **uni-label** que **multi-label**... ;
- **ID91** : La conférence EGC'2017 propose un défi dont le contexte est la gestion des espaces verts pour la ville de Grenoble... rappel de 0,91 sur la prédiction **unilabel**... 0,46 pour la prédiction **multilabel** ;
- **ID361** : Sélection d'une méthode de classification **multi-label** pour un système interactif.

Troisième groupe : les systèmes de recommandation.

- **ID255** : Extraction de l'intérêt implicite des utilisateurs dans les attributs des items pour améliorer les **systèmes de recommandations**. Les **systèmes de recommandation** ont pour objectif... ;
- **ID271** : Pour une meilleure exploitation de la classification croisée dans les systèmes de filtrage collaboratif : ... meilleure exploitation du potentiel de la classification croisée dans le domaine des **systèmes de recommandation**... ;
- **ID121** : Recommandations et prédictions de préférences basées sur la combinaison de données sémantiques et de folksonomie : Dans les **systèmes de recommandation**,

l'approche du filtrage... ;
— **ID77** : Application mobile pour l'évaluation d'un algorithme de calcul de distance entre des items musicaux : Les **systèmes de recommandation** permettent de présenter... ;
— **ID197** : PersoRec : un **système** personnalisé de **recommandations** pour les folksonomies basé sur les concepts quadratiques ;
— **ID3** : ALGeoSPF : Un modèle de factorisation basé sur du clustering géographique pour la **recommandation** de POI : La **recommandation** de points d'intérêts... ;
— **ID184** : Intégration des Influences Géographique et Temporelle pour la **Recommandation** de Points d'Intérêt : La **recommandation** de points d'intérêts (ou POI), est devenue... ;
— **ID38** : L'exploitation de données contextuelles pour la **recommandation** d'hôtels : Les **systèmes de recommandation** ont pour rôle d'aider... ;
— **ID423** : Technique de factorisation multi-biais pour des **recommandations dynamiques** : La factorisation de matrices offre une grande qualité de prédiction pour les **systèmes de recommandation**...

L'article d'ID 241, paru en 2015 aborde même les deux thématiques : « Cet article présente une solution centrée sur les ontologies pour la classification **multi-label** automatique d'information nécessaire à un **système de recommandation** d'informations économiques. »

4.4 Analyse des résultats

Il est assez difficile de juger la qualité des résultats obtenus, puisse qu'il s'agissait d'un défi ouvert, mais on peut, tout de même, faire les observations suivantes :

— Le mot « label » apparaît dans 10 documents seulement : 1 fois en 2014, 2 fois en 2015 et 7 fois à partir de 2016. Il est associé 8 fois au mot « multi ». On peut donc considérer que la thématique « **multi-label** » est réellement nouvelle à partir de 2016.
— Si le sujet de la « **recommandation** » existait déjà dans le passé, on constate dans les données une forte augmentation de cette thématique. Ce mot est mentionné dans 35 documents, 16 fois entre 2004 et 2014 et 19 fois entre 2015 et 2018, soit plus de fois en 4 ans qu'en 11 ans, ce qui traduit bien son caractère nouveau ou de re-nouveau.

5 Conclusion

Le précédent défi d'EGC sur l'analyse des thématiques et de leurs évolutions datant d'il y a seulement 3 ans, nous avons cherché à détecter les thématiques nouvelles ou émergentes ces dernières années. Pour cela, nous avons utilisé une technique de visualisation basée sur les graphes et une approche de type « word embeddings » et « document embeddings ». Au final, ces techniques nous ont permis de détecter deux thématiques récentes dans la communauté EGC : la **classification multi-label** et les **systèmes de recommandation** (de musique, de lieux, d'hôtels, etc.).

Références

Bastian, M., S. Heymann, et M. Jacomy (2009). Gephi : an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media*.

Cabanac, G., G. Hubert, H. D. Tran, C. Favre, et C. Labbé (2016). Un regard lexico-scientométrique sur le défi egc 2016.

Guille, A., E.-P. Soriano-Morales, et C.-O. Truica (2016). Topic modeling and hypergraph mining to analyze the egc conference history. In *EGC*, pp. 383–394.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, et J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Németh, L., V. Trón, P. Halácsy, A. Kornai, A. Rung, et I. Szakadát (2004). Leveraging the open source ispell codebase for minority language analysis. *Proceedings of SALTMIL*, 56–59.

Page, L., S. Brin, R. Motwani, et T. Winograd (1999). The pagerank citation ranking : Bringing order to the web. Technical report, Stanford InfoLab.

Ratinaud, P. (2009). IRaMuTeQ : Interface de R pour les analyses Multidimensionnelles de Textes et de Questionnaires, `http://www.iramuteq.org`.

Reinert, M. (1983). Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Cahiers de l'Analyse des Données 8*(2), 187–198.

Reinert, M. (1986). Un logiciel d'analyse lexicale. *Cahiers de l'analyse des données 11*(4), 471–481.

Shen, D., G. Wang, W. Wang, M. R. Min, Q. Su, Y. Zhang, C. Li, R. Henao, et L. Carin (2018). Baseline needs more love : On simple word-embedding-based models and associated pooling mechanisms. *arXiv preprint arXiv :1805.09843*.

Suignard, P., M. Bothua, et A. Benamar (2019). Participation d'EDF R&D à DEFT 2019 : des vecteurs et des règles ! *DEFT*.

Torki, M. (2018). A document descriptor using covariance of word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pp. 527–532.

Summary

In this article, we present how word and document embeddings techniques, learned on the corpus of EGC conference articles, associated with graph visualization, allow us to detect emerging themes of this conference.