Analyses synchroniques et diachroniques des thématiques EGC- Défi ECG 2016

Sofiane Bouzid*, Adrian Tanasescu*

*Institut des Sciences de l'Homme, Lyon
sofiane.bouzid@ish-lyon.cnrs.fr
adrian.tanasescu@ish-lyon.cnrs.fr

Résumé. Les articles scientifiques publiés dans les actes des conférences EGC, qui se déroulent chaque année depuis 2001, constituent la richesse de ces évènements mettant en avant le fer de lance de la recherche francophone portant sur la gestion et l'extraction de connaissances. Nous nous sommes penchés sur l'analyse de ces publications scientifiques afin d'en extraire l'essence en termes de thématiques de recherches abordées. Premièrement, nous avons analysé les points communs et les spécificités des publications dans les différentes éditions de la conférence ainsi que les principales différences entre les éditions consécutives. Puis nous nous sommes intéressés à la façon dont les publications s'articulent autour des thématiques extraites et sur lesquelles nous avons essayé de visualiser une approximation sémantique. Enfin nous nous sommes intéressé à l'évolution des thématiques depuis les débuts de cette conférence et jusqu'à l'édition 2015.

1 Introduction et préliminaires

La littérature scientifique se distingue par une structuration qui diffère selon la discipline de l'article. Des unités essentielles, appelées les clés de texte (titre, résumé, mots-clés), sont placées au début de chaque articles, elles permettent de décrire d'une façon brève et efficace tout son contenu (Bénichoux et al. (1985)).

Les analyses présentées dans cet article, ont porté sur le contenu textuel décrit dans les titres et les résumés des publications scientifiques présentées lors des 12 dernières éditions de la conférence EGC.

Premièrement, nous nous sommes concentrés sur l'analyse des occurrences des termes puis la visualisation des fréquences des termes les plus fréquents qui permettent de ressortir les orientations des articles d'une façon orientée *mots-clés*.

Puis, nous avons utilisé des algorithmes de détection de thématiques afin de détecter des typologies de publications.

Enfin, nous avons étudié l'évolution des thématiques découvertes au fur et à mesure des éditions successives de la conférence.

2 Description du corpus de données

Afin de pouvoir analyser les orientations et thématiques de ces articles, nous avons isolé ceux dont la langue de rédaction était le français. Pour cela nous avons utilisé un algorithme de détection de la langue de textes basé sur les profiles n-gram proposé dans Hornik et al. (2013).

Sur l'ensemble de la période 2004-2015 les articles en français représentent environ 88 % du nombre total de publications EGC.

3 Formalisation du problème

Pour analyser le contenu textuel des articles scientifiques, nous utilisons la représentation vectorielle décrite dans Salton et al. (1975). Cela consiste à représenter un corpus par une matrice où les lignes représentent des descripteurs et les colonnes les documents. Une cellule de la matrice contiendra la fréquence d'apparition d'un terme dans un document.

4 Orientation des articles de recherche selon les éditions

Cette méthode d'analyse orientée *mots-clés* nous a permis d'obtenir une représentation globale des orientations des articles scientifiques selon les éditions d'EGC entre 2004 et 2015.

Nous avons souhaité suivre l'évolution de ces termes en mettant en évidence des orientations qui apparaissent, disparaissent et qui sont constantes d'une édition à la suivante.

Ainsi, les représentations visuelles dans le figure 1 mettent en évidence les principaux termes communs sous forme de disques bleus, les principaux termes qui sont apparus dans l'édition suivante en vert, enfin les principaux termes ayant disparus depuis l'édition précédente en rouge.

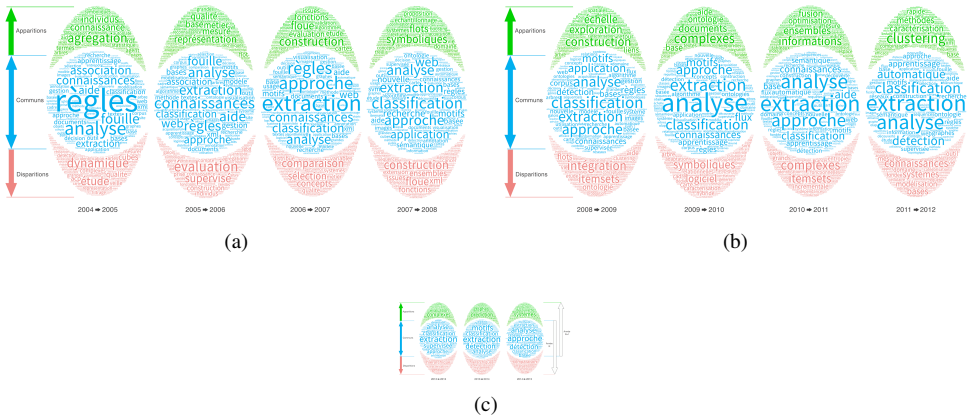
Bien évidemment, dans la lecture de ces visualisations, les termes communs et les nouveaux termes définissent l'année *N+1* alors que les termes communs et les termes disparus décrivent l'année *N*.

Les interprétations de lecture de ces représentations visuelles sont nombreuses. On peut, par exemple, observer la persistance des termes comme *analyse, classification, extraction et apprentissage* dans l'ensemble des éditions depuis 2004. On peut également observer, par exemple, l'apparition des termes *agrégation, arbres et statistique* lors de l'édition 2005, termes qui n'étaient présents dans aucun article en 2004 lors de la première édition de la conférence.

Les termes *qualité, représentation* et *owl* font leur apparition en 2006, alors que le termes *évaluation* et *supervisé* n'apparaissent plus dans les articles de cette même année. L'année 2008 voit apparaître des articles portant sur la *symbolique* et les *treillis et concepts* et constate une disparition des articles sur le sujet des *ensembles flous*.

On peut également observer, par exemple, une disparition des termes *similarité* et *hiérarchique* lors de l'édition 2014, édition qui fait réapparaître les termes *graphes* et *prédiction*.

Enfin, nous avons extrait les termes communs entre les éditions d'EGC. On remarque que les termes tels que *extraction, classification, analyse, apprentissage, fouille, visualisation et règles d'association* définissent en quelque sorte l'essence de la conférence puisqu'ils pourraient résumer à eux seuls le domaine d'EGC.



(a) (b)

(c)

FIG. 1: Évolution des termes entre les éditions successives d'EGC entre 2004 et 2015

5 Détection et analyse des thématiques

L'objectif ici est de découvrir les thématiques latentes dans le corpus des publications, afin d'identifier les axes de recherches d'EGC. Pour ce faire, nous avons utilisé le modèle probabiliste LDA , Latent Dirichlet Allocation (Blei et al. (2003)). Il s'agit d'une modélisation qui fournit un puissant outil pour découvrir et exploiter la structure thématique cachée dans des corpus textuels (Blei (2012)).

Nous avons utilisé le corpus englobant la totalité des articles scientifiques durant la période d'observation pour l'extraction des thématiques. Préalablement, nous avons procédé à un ensemble de pré-traitements (omission des mots-outils, des ponctuations, etc.) tout en gardant les termes rares afin de ne pas éliminer les éventuelles spécificités. Nous avons également choisi de ne pas effectuer de lémmatisation ou de racinisation (*stemming*) du vocabulaire car le corpus est très spécialisé.

Nous avons testé différents nombres de thématiques pouvant émerger, nous avons retenu le regroupement en 10 thématiques. La pertinence et la cohérence des thématiques ont été déterminées en analysant les termes selon leur fréquence dans le corpus et leurs probabilité d'appartenance à une thématique donnée (Sievert et Shirley (2014)). Les tendances qui se dégagent nous paraissent représentatives des articles publiés dans les conférences EGC. Les dix thématiques retenues sont :

Thématique 1 : *Ontologies, sémantique et annotation de corpus de documents* ; Thématique 2 : *Représentations et explorations visuelles, génétique* ; Thématique 3 : *Règles et extraction de motifs fréquents* ; Thématique 4 : *Traitement d'images/vidéos et séquences spatio-temporelles* ; Thématique 5 : *Représentation de concepts, symbolique et sémantique* ; Thématique 6 : *Entrepôts de données et analyse multidimensionnelle* ; Thématique 7 : *Partitionnement et cartographie, clustering* ; Thématique 8 : *Méthodes d'apprentissage supervisé, classification, arbres* ; Thématique 9 : *Graphes et réseaux de communautés* ; Thématique 10 : *Recherche d'information, corpus textuels et documents XML* ;

5.1 Visualisation des thématiques et proximité sémantique

Nous avons effectué une analyse de positionnement multidimensionnel (*Multidimensional scaling*) sur la matrice de dissimilarité entre thématiques construite en utilisant la mesure *divergence* de Kullback-Leibler. Dans notre cas, le positionnement MDS permet de cartographier les thématiques et d'avoir une approximation de la proximité sémantique entre les thématiques. La figure 2 permet de visualiser les thématiques, sur le plan engendré, décrites par les nuages de mots-clés les plus représentatifs. La taille du nuage de termes de chaque thématique est relative au nombre de documents dans la thématique et la taille des termes dans le nuage des thématiques est proportionnelle à leur fréquence mais aussi à leur spécificité. Ainsi, un terme très fréquent et dont la plupart des occurrences est parmi les documents d'une thématique sera un mot plus grand dans le nuage de cette thématique.

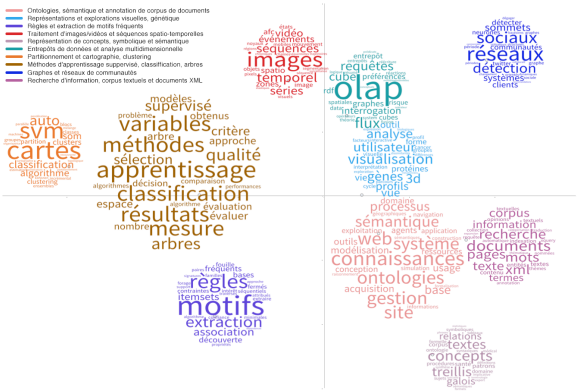


FIG. 2: Carte des distances inter-thématiques MDS

Généralement on estime l'appartenance d'un terme à une thématique en utilisant une probabilité d'appartenance. (Sievert et Shirley (2014)) proposent une nouvelle approche dans ce sens, basée sur la pondération à l'aide d'un paramètre λ compris entre 0 et 1 qui représente la distribution estimée du terme dans les documents de la thématique par rapport à la distribution marginale du terme dans le corpus global. L'utilisation de cette pondération à une valeur proche de 0.5, nous permet d'obtenir les termes spécifiques de chacune des thématiques (en sélectionnant ainsi ceux dont la majorité des occurrences se trouvent dans les articles de la thématique).

On peut observer dans la figure 2 que la thématique la plus importante, regroupant le plus d'articles, est celle liée aux *Méthodes d'apprentissage supervisés, classification, arbres*. Deux autres thématiques se retrouvent assez proches sur la carte des distances MDS : la première regroupe les approches basées le *clustering et cartographie* (de Kohonen) alors que la seconde est orientée sur les aspects liés à l'extraction de *Règles et de motifs fréquents*. Nous remarquons aussi la proximité entre la thématique *Entrepôts de données et analyse multidimensionnelle* et la thématique liée au *Traitement d'image/vidéo et séquences spatio-temporelles"* et enfin de la thématique liée aux *Représentations, explorations visuelles et à l'analyse de données génétiques*. On note également une proximité non négligeable (dans le quart en haut et à droite de la

représentation MDS) entre la thématique qui traite des *"Graphes et réseaux de communautés"* et les deux thématiques précédentes ("OLAP" et "visualisation").

On peut observer l'existence de trois thématiques concentrées dans le quart en bas et à droite du plan de projection, ce qui dénote une proximité sémantique entre elles. Il s'agit des thématiques relatives aux *Ontologies, sémantique et annotation de corpus de documents* (la plus représentée des trois), à la *Recherche d'information, corpus textuels et documents XML* et enfin à la *Représentation de concepts, symbolique et sémantique*. Cette proximité semble naturelle car il s'agit d'approches s'intéressant aux documents textuels structurés ou non ainsi qu'à leur sémantique.

5.2 Diachronie des thématiques

Afin de visualiser l'évolution des thématiques détectées précédemment, nous avons effectué une analyse du nombre de documents dans les thématiques au fur et à mesure des 12 éditions d'EGC.

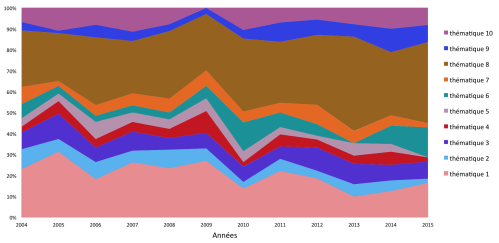


FIG. 3: Évolution des % de représentations des thématiques entre les éditions d'EGC

La figure 3 présente ces évolutions sous forme d'un graphique en aires permettant une mise en lumière de l'amplitude de la variation des thématiques en pourcentage de l'ensemble des publications par année. La figure 3 montre que la thématique *Méthodes d'apprentissage supervisé, classification, arbres* et la thématique *Ontologies, sémantique et annotation de corpus de documents* sont deux grands axes de recherche qui alimentent presque 2/3 des articles de la revue et cela durant l'ensemble de la période 2004-2015.

On note qu'en 2009 il y a eu une baisse significative concernant la thématique 10 liée à la *Recherche d'information, corpus textuels et documents XML"*, tout comme en 2013 la thématique 6 liée aux *Entrepôts de données et à l'analyse multidimensionnelle* a été absente pour revenir fortement en 2015. Cette même dernière année, la thématique 5 liée à la *Représentation de concepts, symbolique et sémantique* a diminué fortement.

Concernant l'évolution de chacune des thématiques nous constatons que :

la thématique 3 *Règles et extraction de motifs fréquents* est bien représentée en 2005 et 2012 ; la thématique 6 *Entrepôts de données et analyse multidimensionnelle* est sensiblement mieux représentée en 2010 et 2015 comparé aux autres éditions ; la thématique 9 *Graphes et réseaux de communautés* connaît un point culminant en 2014 ; la thématique 2 *Représentations et explorations visuelles, génétique* est bien représentée en 2004 et 2008 ; la thématique *Ontologies, sémantique et annotation de corpus de documents* connaît ponctuellement une

baisse significative en 2010. On observe que ce sont les articles de cette même thématique qui présentent une réelle tendance, baissière, entre 2004 et 2015.

6 Conclusion

Nous avons essayé dans ce travail de faire émerger et de visualiser les axes et orientations des articles scientifiques publiés dans les actes des conférences EGC. Ce travail a été mené selon deux approches, une empirique, basée sur le calcul des occurrences des termes, une autre probabiliste, basée sur les co-occurrence des termes et leurs distributions de probabilité. Cela nous a permis de découvrir des thématiques autour desquelles s'articulent les articles scientifiques publiés durant la période observée de 12 ans. Dans la dernière partie, nous avons essayé de visualiser une similarité sémantique entre les thématiques puis d'observer leur évolution entre 2004 et 2015. Il serait intéressant de prolonger ces analyses pour voir si les changements de thématiques étaient liés à l'arrivée de nouveaux auteurs dans la communauté EGC et également d'envisager la constitution d'un réseau social relativement aux thématiques détectées.

Références

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM 55*(4), 77–84.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *J. Mach. Learn. Res. 3*, 993–1022.

Bénichoux, R., D. Pajaud, et J. Michel (1985). *Guide pratique de la communication scientifique*. Paris : G. Lachurié.

Hornik, K., P. Mair, J. Rauch, W. Geiger, C. Buchta, et I. Feinerer (2013). The textcat package for n-gram based text categorization in R. *Journal of Statistical Software 52*(6), 1–17.

Salton, G., A. Wong, et C. S. Yang (1975). A vector space model for automatic indexing. *Commun. ACM 18*(11), 613–620.

Sievert, C. et K. E. Shirley (2014). LDAvis : A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pp. 63–70.

Summary

Scientific papers published in the proceedings of EGC, conference that is held annually since 2001, constitute the richness of these events that put in scene the best of the French research focusing on knowledge extraction and management. In this paper we analyzed the scientific papers of the last 12 editions of EGC in order to highlight the main research subjects that were addresses. First, we detected common terms and specificities between the articles of different editions of the conference and also differences that could appear between consecutive editions. Then we studied how articles can be grouped into separate topics, that we automatically detected, and how these topics evolved since the conference beginnings until 2015.