Un regard lexico-scientométrique sur le défi EGC 2016

Guillaume Cabanac*, Gilles Hubert*, Hong Diep Tran*, Cécile Favre**, Cyril Labbé***

*Université de Toulouse, IRIT UMR 5505
**Université de Lyon, ERIC ÉA 3083
***Université de Grenoble, LIG UMR 5217

Résumé. Depuis 2001, les conférences EGC ont rassemblé 1 782 chercheurs autour de l'extraction et la gestion de connaissances. En 2016, l'association EGC réfléchit à son histoire et se projette en lançant un défi à sa communauté. Que peut-on révéler sur la communauté EGC *via* des approches développées en EGC ? Notre étude lexico-scientométrique apporte un éclairage sur les thématiques du congrès, les lieux de publication investis par ses auteurs, ou encore les auteurs sollicitables comme évaluateurs. Les résultats sont intégrés à un site web sous-tendu par un système d'information décisionnel.

1 Introduction

Depuis 2001, le congrès EGC rassemble la communauté scientifique autour de l'extraction et de la gestion des connaissances. À l'instar de communautés plus anciennes, EGC se questionne sur son parcours et son futur en lançant le défi « Communauté EGC : quelle histoire et quel avenir ? ». Des initiatives réflexives analogues ont notamment stimulé les acteurs des communautés INFORSID (Collectif INFORSID, 2012) et CAiSE (Jarke et al., 2013).

Nous avons appréhendé le défi EGC comme l'expression d'un besoin en système d'information décisionnel (c.-à-d., facilitant la prise de décision). Notre contribution repose sur l'analyse et la conception d'une solution au service de différents acteurs : la gouvernance d'EGC pour bénéficier de synthèses utiles au pilotage de l'association et à l'organisation des congrès ; les scientifiques préoccupés par les thématiques d'EGC pour identifier des sources ou des cibles de publication ; les scientifiques distants des thématiques d'EGC pour comprendre sur quels objets travaille la communauté et où se situe le front de recherche, par exemple. Sur la base de ces besoins, nous avons conçu et développé une plateforme logicielle en ligne (`http://www.irit.fr/~Guillaume.Cabanac/egc`) visant à valoriser la communauté EGC au travers des fonctionnalités suivantes :

— la *navigation* dans un catalogue hypertextuel de notices bibliographiques structuré par édition du congrès ainsi que la *recherche* d'auteurs et de publications par mots-clés ;
— la *recommandation* de membres de comité de programme selon différents critères ;
— le *positionnement* de la communauté EGC en termes de localisation géographique des éditions du congrès, de thèmes traités dans les articles publiés, de cibles de publication des auteurs, d'analyse sexuée sur la place des femmes dans la communauté EGC.

Notre contribution s'appuie sur les données que nous présentons dans la section suivante, en partie fournies par l'association EGC et en partie extraites par nos soins.

2 Données et méthodes

Notre système d'information décisionnel intègre des données issues de différentes sources. L'association EGC a fourni les métadonnées des 1 061 articles publiés à partir de 2004. Chaque notice bibliographique d'article comprend son titre, ses auteurs, son année de publication et parfois même son résumé. Ces données fournies ont été consolidées et complétées en :

— interrogeant la base bibliographique DBLP (`http://dblp.uni-trier.de/xml`) pour acquérir les 151 notices manquantes des congrès de 2001 à 2003, afin de présenter des résultats couvrant l'intégralité des éditions du congrès ;
— saisissant l'identité des 333 membres de comités de programme (CP) de tous les congrès EGC à partir des actes édités ;
— identifiant le sexe des 1 881 auteurs et membres de comités de programme. Cet important travail manuel nécessitait une connaissance des auteurs d'EGC ou, à défaut, une recherche sur leurs pages Web (ex. photo), des biographies (pronom « il » ou « elle »), des dictionnaires de prénoms, etc. Une méthode automatique n'aurait pu fournir les données avec autant de précision, là où nous avons cherché à établir un bilan chiffré au plus juste de la réalité. Ceci participe à la difficulté de mener des études sexuées à plus grande échelle sur ce type de données sans disposer de données externes fiables ;
— interrogeant DBLP pour acquérir les biographies de ces 1 881 auteurs et membres de comités de programmes, afin de déterminer leurs « lieux de publications » connexes.

Nous avons mobilisé plusieurs méthodes pour exploiter ces données intégrées dans une base de données relationnelles. L'interrogation de cette base permet de construire le catalogue hypertextuel, d'extraire des statistiques descriptives, de confronter les lieux de publications de la communauté EGC issus des données EGC et ceux issus de DBLP. Par ailleurs, la recommandation de membres de CP est réalisée par requête, en exprimant différents critères de sélection et de tri. Enfin, les analyses de thématiques à partir des titres et résumés sont réalisées en extrayant les mots représentatifs du corpus et en appliquant des techniques de lexicométrie.

3 Résultats

Cette section présente les résultats de notre analyse lexico-scientométrique (consultables sur `http://www.irit.fr/~Guillaume.Cabanac/egc`) intégrés au système d'information décisionnel développé. Ils s'ajoutent au catalogue hypertextuel intégrant la liste des articles publiés et des membres du CP de chaque édition ainsi que la carte des éditions d'EGC qui souligne la répartition des promoteurs d'EGC sur le territoire francophone.

3.1 Moteur de recommandation de membres de comité de programme

L'association EGC édite des actes comprenant les articles évalués par les pairs, membres d'un comité de programme usuellement constitué par un président sur la base de divers critères non formalisés. Ces membres sont des chercheurs actifs publiant à EGC ou dans des lieux de publications connexes, et qui représentent le large éventail des thématiques du congrès et la variété des centres universitaires. Sur le temps long, le renouvellement des membres de CP est souhaitable afin d'éviter des situations de concentration de pouvoir et de verrouillage qui sont à l'encontre de l'éthos des sciences (Merton, 1942).

Or, à notre connaissance et d'après des échanges informels avec d'ex-présidents, la constitution de CP repose principalement sur les perceptions et les accointances du président du CP. Bien qu'actifs sur des thématiques du congrès, certains chercheurs, qui pourraient être disponibles, ne sont jamais sollicités car non visibles sur « les radars » des présidents de CP. Il nous a semblé judicieux de tirer parti de l'historique des CP combiné à l'historique des publications pour identifier de tels scientifiques. Nous avons dégagé quatre indicateurs à partir desquels réaliser la recommandation : la dernière édition d'EGC à laquelle le chercheur a participé comme membre de CP, le nombre d'articles publiés à EGC jusqu'alors, la dernière édition d'EGC dans laquelle le chercheur a publié, et enfin le nombre de participations comme membre de CP.

Ces indicateurs sont employés comme clés de tri des scientifiques recensés dans notre base de données. L'implémentation actuelle promeut les scientifiques qui n'ont jamais participé à des CP EGC, qui sont les plus présents en termes de publications en favorisant d'abord ceux ayant *récemment* publié à EGC puis les moins sollicités pour de précédents CP.

Conscients de la nature arbitraire des critères retenus, il sera judicieux de réviser ces choix à la lumière de deux études à réaliser. D'une part, il s'agira de questionner un panel de présidents de CP, pas uniquement d'EGC, pour éliciter l'ensemble des critères qu'ils utilisent communément. D'autre part, il s'agira d'analyser la composition d'une variété de CP, pas uniquement d'EGC, pour inférer des critères de sélection et leur combinaison.

3.2 Positionnement d'EGC

Les différents acteurs intéressés par les activités d'EGC peuvent chercher à connaître les thèmes de recherche investis et les autres cibles de publication des auteurs. Nous avons mobilisé des techniques de lexicométrie et d'extraction d'information pour satisfaire ce besoin.

3.2.1 Thématiques des congrès EGC

Afin de distinguer les courants principaux des congrès EGC, nous avons réalisé une classification des titres des articles par une approche phare de lexicométrie : la méthode de Reinert (1983) implémentée dans la plateforme Iramuteq (Ratinaud, 2009). Deux blocs principaux se dégagent de cette classification descendante non-supervisée (figure 1). D'une part, les classes 4 et 3 illustrent des travaux liés aux processus (acquisition, gestion, accès, recherche, capitalisation, etc.) et aux acteurs (utilisateur, agent, organisation). D'autre part, les classes 1, 2 et 5 regroupent des travaux plus techniques et liés aux méthodes (apprentissage, clustering, reconnaissance) et aux objets de recherche (cube, graphe, motif, règle, séquence, etc.).

L'analyse automatique des mots reliés et leur visualisation graphique (voir en ligne) proposés par Iramuteq révèle les concepts d'intérêt au sein de la communauté EGC, comme par exemple, les concepts d'« extraction de motifs fréquents/séquentiels », « analyse OLAP » et « visualisation interactive de données ».

3.2.2 Caractérisation des thèmes par l'étude des champs lexicaux

La compréhension profonde des thèmes abordés par un corpus nécessite de cerner le sens du vocabulaire employé. On considère généralement que le sens d'un vocable dépend de ceux qui l'entourent (*amour filial, amour de la patrie, etc.*). Pour nombre de tâches en TALN, un vocable se caractérise par les fréquences observées dans son entourage (Pennington et al., 2014).

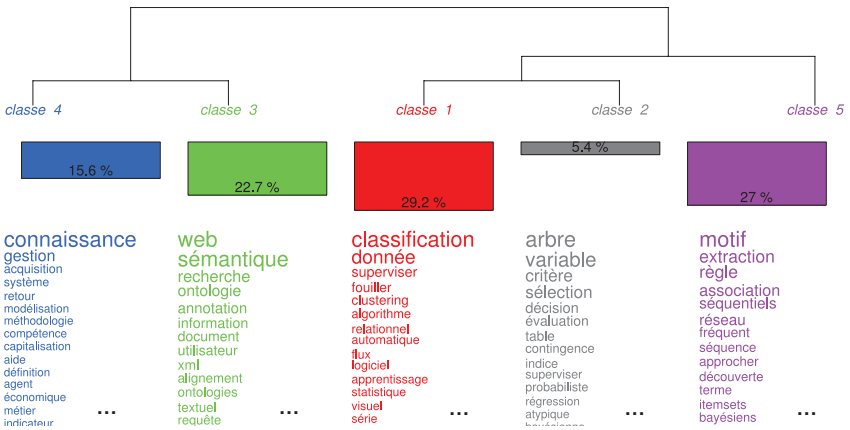

FIG. 1 – *Classification des titres des 1 212 publications EGC.*

Pour la compréhension fine des concepts utilisés dans un corpus, il est tout aussi important de considérer les vocables qui sont sur-représentés ou sous-représentés autour de ce concept (Labbé et Labbé, 2005). L'étude du vocabulaire sur/sous-employé dans l'univers de *données* dans le corpus EGC montre que *données* est plus souvent associé à *fouille* que à *web* ou *internet*. La même étude pour *connaissance* révèle que, dans le corpus EGC, *connaissance* est plus fortement associé à *extraire, extraction* qu'à *classification* et *classer*.

3.2.3 Lieux de publication investis par les auteurs d'EGC

Dans cette section, nous tentons d'apporter une réponse à des questions récurrentes formulées par les acteurs impliqués dans la communauté vis-à-vis des conférences et revues liées à l'EGC. Les *lecteurs* se demandent où cibler leur veille scientifique et technique, les *auteurs* se demandent où diffuser leurs contributions scientifiques, et enfin, la *gouvernance* d'EGC se demande certainement à qui proposer des actions conjointes (numéros spéciaux de revues ou co-éditions de congrès) pour rapprocher les communautés.

Il n'existe pas, à notre connaissance, une telle ressource indiquant les liens tissés entre les communautés. Nous proposons d'exploiter les bibliographies des chercheurs d'EGC pour identifier où ils publient. À partir de DBLP, nous avons extrait les lieux de publications privilégiés par les auteurs d'EGC en termes de conférences et de revues, parmi lesquelles nous retrouvons outre EGC, DEXA, BDA, ICPR ou encore ICASSP. Une présence simultanée de lieux francophones et internationaux suggère que les auteurs d'EGC sont actifs sur la scène internationale tout en étant attachés à la diffusion de leurs travaux au niveau national.

3.2.4 Les femmes dans la communauté EGC

La question de la place des femmes, aujourd'hui transversale dans la société, n'échappe pas au domaine de la recherche, notamment au travers des questionnements sur l'égalité professionnelle. Ainsi, au travers de ce regard scientométrique dans le cadre du défi EGC, compte tenu du constat, qu'en France, les études en informatique sont de plus en plus délaissées

par les femmes (constats partagés lors du Congrès Femmes et Informatique de 2015 [1]), il nous a paru pertinent d'observer la communauté EGC à cet égard.

Concernant la présidence du CP, deux femmes ont exercé ce rôle sur l'ensemble des éditions et ce, de façon récente (2014, 2015). Concernant la participation au CP et la contribution au travers des articles, nous mettons en perspective les chiffres d'EGC avec des statistiques plus globales fournies par le ministère en 2012 sur le « vivier » d'enseignants-chercheurs (pour les sections CNU 27 et 26 concernant la communauté EGC on dénombrait respectivement, environ 80 à 85 % d'hommes PR, et 75 à 80 % pour les MCF). Concernant les contributions scientifiques, au prorata du nombre d'auteurs dans les articles, globalement entre 2001 et 2015, il existe un rapport de 26 % de femmes pour 74 % d'hommes, avec une oscillation entre 21 % en 2011 et 30 % à plusieurs reprises dont 2012. Ce résultat montre que, globalement, la communauté des auteurs d'EGC, bien que très masculine, se situe dans les chiffres officiels du ministère, voire légèrement plus féminine. Les stéréotypes de genre peuvant accentuer la présence ou l'absence des femmes selon les champs de l'informatique, il serait intéressant d'étudier les différentes communautés (systèmes multi-agent, réseau, systèmes d'information, etc.) de façon comparative.

En étude sur le genre, un indicateur quantitatif utilisé est le rapport de masculinité, exprimé en nombre d'hommes pour 100 femmes. À la naissance, le rapport de masculinité est dans la plupart des pays de 105 garçons pour 100 filles. En calculant ce rapport sur les membres du CP, après un démarrage de la conférence avec un CP très masculin (7 femmes pour 28 hommes, soit un rapport de masculinité de 536 %), le CP s'est ensuite féminisé pour atteindre des rapports de 250 à 280 % selon les années, soit 72 à 75 % d'hommes. Nous constatons néanmoins une tendance à la hausse de ce rapport de masculinité depuis 2003, ce qui questionne quant aux critères de constitution des CP.

4 Travaux connexes : la scientométrie pour étudier la science

Étudier la science nécessite de modéliser, analyser et visualiser les acteurs de la recherche et de l'innovation ainsi que leur production. L'*atlas des sciences* de Börner (2010) offre un panorama des approches développées depuis la renaissance à cet effet et, plus récemment, en *scientométrie*, champ scientifique désignant l'étude quantitative de la science et de l'innovation (Leydesdorff et Milojević, 2015). Parmi les approches, on trouve principalement les réseaux de coauteurs, de références et de thématiques, qui sous-tendent, notamment, la recherche d'experts (Balog et al., 2012) ou l'estimation de la qualité de conférences (Zhuang et al., 2007).

5 Conclusion et perspectives

Les techniques lexico-scientométriques mobilisées dans cet article ont élicité les thématiques du congrès EGC, les lieux de publication des membres de la communauté ainsi que les scientifiques sollicitables dans l'optique de futurs CP. Ces éléments s'articulent au sein d'un système d'information décisionnel conçu comme une vitrine en ligne de l'association EGC. Pour prolonger cette première réalisation, il semble pertinent de caractériser le front de recherche à l'international (par ex., poussée de l'apprentissage profond) pour raffiner la portée

1. http://www.univ-orleans.fr/lifo/evenements/SIF2015

des futurs appels à communications. En outre, l'analyse des réseaux de citations révélerait les auteurs et thématiques plébiscités par la communauté EGC ; ces derniers pourraient alors être sollicités dans le cadre, notamment, de conférences invitées lors des congrès EGC.

Références

Balog, K., Y. Fang, M. de Rijke, P. Serdyukov, et L. Si (2012). Expertise retrieval. *Foundations and Trends in Information Retrieval 6*(2–3), 127–256.

Börner, K. (2010). *Atlas of Science : Visualizing What We Know*. Cambridge, MA : MIT Press.

Collectif INFORSID (2012). La recherche en systèmes d'information et ses nouvelles frontières. *Ingénierie des Systèmes d'Information 17*(3), 9–68.

Jarke, M., M. C. Pham, et R. Klamma (2013). Evolution of the CAiSE author community : A social network analysis. In *Seminal Contributions to Information Systems Engineering*, pp. 15–33. Berlin : Springer.

Labbé, C. et D. Labbé (2005). How to measure the meanings of words ? Amour in Corneille's work. *Language Resources and Evaluation 39*(4), 335–351.

Leydesdorff, L. et S. Milojević (2015). Scientometrics. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2 ed.), Volume 21, pp. 322–327. Elsevier.

Merton, R. K. (1942). Science and technology in a democratic order. *Journal of Legal and Political Sociology 1*(1), 115–126.

Pennington, J., R. Socher, et C. D. Manning (2014). Glove : Global vectors for word representation. In *EMNLP'04 : Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, pp. 1532–1543. ACL.

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. *Les cahiers de l'analyse des données 8*(2), 197–198.

Zhuang, Z., E. Elmacioglu, D. Lee, et C. L. Giles (2007). Measuring conference quality by mining program committee characteristics. In *JCDL'07 : Proceedings of the joint conference on digital libraries*, New York, NY, pp. 225–234. ACM.

Summary

The French-speaking EGC congress gathered scholars since 2001 to tackle issues on knowledge discovery and management. EGC questions its history and future in the « défi EGC » challenge. Our lexico-scientometric approach sheds the light on the topics and related venues favourited by its contributing authors. We identify those active scholars who are seldom invited as PC members by relating data from the program committees and authorships. Our results are featured in a decision support system showcasing the activities of the EGC association.