Il était une fois Algotel

Fabien Mathieu, Sébastien Tixeuil

Il était une fois Algotel †

Fabien Mathieu[1] et Sébastien Tixeuil[2]

[1]*Swapcard, Paris, France*
[2]*Sorbonne Université, CNRS, LIP6*

Les notions de « communauté scientifique » et de « domaine de recherche » sont des éléments centraux pour les chercheurs et les articles qu'ils publient. Nous proposons d'explorer l'évolution de la communauté Algotel depuis sa création à partir des articles recensés dans DBLP et des comités de programme. Nos résultats permettent d'une part de mieux comprendre l'évolution de la communauté, et d'autre part d'intégrer facilement de nouvelles thématiques ou chercheurs lors des éditions futures.

Mots-clefs : Bibliométrie, Communauté, Domaine scientifique, Algotel

1 Introduction

Comprendre les progrès de la science en étudiant comment de nouvelles connaissances scientifiques sont créées constitue un enjeu sociétal important. Il existe en particulier de nombreuses études concernant la manière dont les connaissances scientifiques se propagent à travers la littérature scientifique. Par exemple, le domaine de la bibliométrie mesure les propriétés du corpus de recherche, et conduit à des estimations d'importance basées sur des notions telles que le nombre de citations d'un article, le facteur d'impact d'une revue et l'indice *h* d'un auteur. Par ailleurs, les aspects sociaux associés à la recherche scientifique, tels que la *sociologie des connaissances scientifiques*, y compris les structures sociales et les processus de l'activité scientifique, ainsi que ses aspects politiques, ont également fait l'objet d'études [4].

Dans cet article, nous nous intéressons à l'analyse de communautés scientifiques, d'un point de vue à la fois spatial et temporel. La dimension spatiale peut être appréciée classiquement via les co-publications des chercheurs [5], mais aussi thématiquement en examinant le vocabulaire utilisé dans les articles publiés. La dimension temporelle rend compte de l'évolution des aspects mesurés au cours du temps. Cette approche est illustrée à travers une étude de la communauté Algotel.

2 Méthodologie

L'originalité de cet article réside dans le fait qu'au lieu de chercher à calculer l'importance supposée d'une communauté ou d'une thématique par rapport à une autre, nous souhaitons caractériser les liens de similarité qui les relient. Pour ce faire, nous employons l'approche résumée dans la Figure 1 :

— À partir de la base de données du *Digital Bibliography & Library Project* (DBLP), nous extrayons un sous-ensemble centré autour de la communauté Algotel ;
— À partir de ce sous-ensemble, nous construisons un double plongement lexical permettant de représenter auteurs et mots dans un même espace, et ainsi de les comparer ;
— En calculant la similarité cosinus entre les vecteurs (l'« angle »), nous pouvons voir les liens entre auteurs, comités de programme, et thèmes.

Nous détaillons maintenant les deux premiers points, le troisième étant l'objet de la section 3.

2.1 Extraction d'un graphe de communauté

Le projet DBLP vise à répertorier les publications anglophones dans le domaine Informatique. La base de données est accessible à l'adresse `https://dblp.uni-trier.de/xml/dblp.xml.gz`, et contient l'ensemble des références bibliographiques, en particulier pour chaque article son titre et ses auteurs.

†Ce travail a été effectué au LINCS (`https://www.lincs.fr/`).

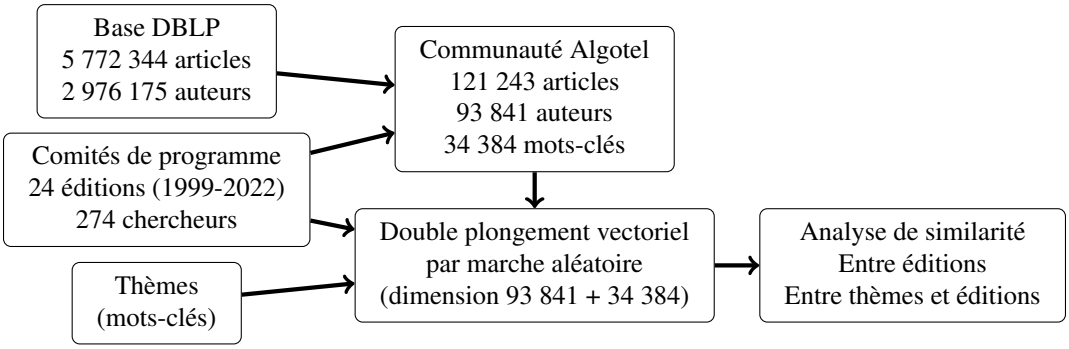


FIGURE 1 – Description synthétique de l'approche

À partir de la base complète, nous avons effectué un filtrage centré sur la communauté Algotel. Pour cela, nous avons collecté l'ensemble des comités de programme des différentes éditions. Pour chaque édition, nous avons effectué une marche aléatoire finie sur un graphe biparti reliant articles et auteurs, à partir des membres du comité de l'édition considérée, et nous avons agrégé les articles les plus probables selon chaque marche (ceux sur lesquels la marche «atterrit» le plus souvent). Cela nous a permis de passer de 5 772 344 articles (taille globale de DBLP) à 121 243 articles issus de la communauté Algotel et de son voisinage.

À partir de ces articles, nous avons construit un graphe biparti reliant 93 841 auteurs à 34 384 mots-clés [‡]. Le graphe relie un auteur et un mot-clé si un auteur a employé le mot-clé dans le titre d'au moins un article.

Les arêtes (dirigées) ont des poids qui prennent en compte les propriétés statistiques des éléments, d'une manière qui généralise le principe *Term Frequency, Inverse Document Frequency* (TF-IDF) : plus un mot est rare, plus il est considéré comme signifiant, et donc plus le poids d'une arête vers ce mot est important [2, 1]. Pour chaque sommet, les arêtes sortantes sont renormalisées pour sommer à 1, ce qui permet de voir le graphe comme une chaîne de Markov.

Cette approche a quelques limitations. D'une part, DBLP ne donne dans sa version téléchargeable aucune information sur le contenu des articles au-delà de leur titre, ce qui peut sembler insuffisant pour effectuer une analyse pertinente. D'autre part, nous n'avons pas pris en compte les années de publication lors de la construction des graphes. En particulier, chaque auteur est analysé à travers l'ensemble de sa carrière, même si ses propres intérêts de recherche ont évolué. Néanmoins, nous verrons que ces limitations n'empêchent pas notre approche de mettre en évidence des tendances.

2.2 Représentation vectorielle

Une marche aléatoire de longueur finie permet de représenter finement un sous-ensemble pondéré de sommets d'une chaîne de Markov par un vecteur de dimension le nombre de sommets [3, 6]. Nous appliquons cette approche sur le graphe de communauté : à tout sous-ensemble (auteur, comité de programme, mot(s)-clé(s)), nous pouvons associer un vecteur sur l'espace des auteurs et des mots du graphe obtenu par une marche aléatoire de longueur géométrique. En regardant l'angle entre deux vecteurs ainsi construits, nous obtenons ainsi une mesure de similarité exploitable.

Implantation et détails techniques : Nos résultats ont été obtenus à l'aide du paquet Python Gismo [2]. Le code utilisé pour cet article est disponible à l'adresse suivante :
`https://github.com/balouf/conference_analysis/tree/main/Algotel`.

3 Résultats

La figure 2 présente la matrice des similarités des communautés scientifiques entre les différentes éditions d'Algotel depuis sa création. Une couleur chaude indique une forte similarité, tandis qu'une couleur froide est le signe d'une faible similarité. On peut faire les observations suivantes. Il existe cinq couples d'années successives à très forte similarité (2001-2002, 2005-2006, 2006-2007, 2013-2014, 2020-2021) qui peuvent

‡. Un mot-clé peut être ici un mot unique, un bi-mots (e.g. *Internet traffic*), ou un tri-mots (e.g. *Wireless sensor networks*).

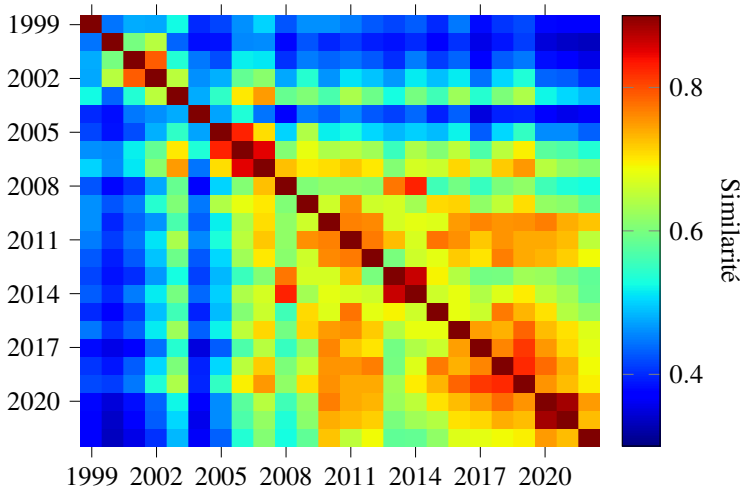


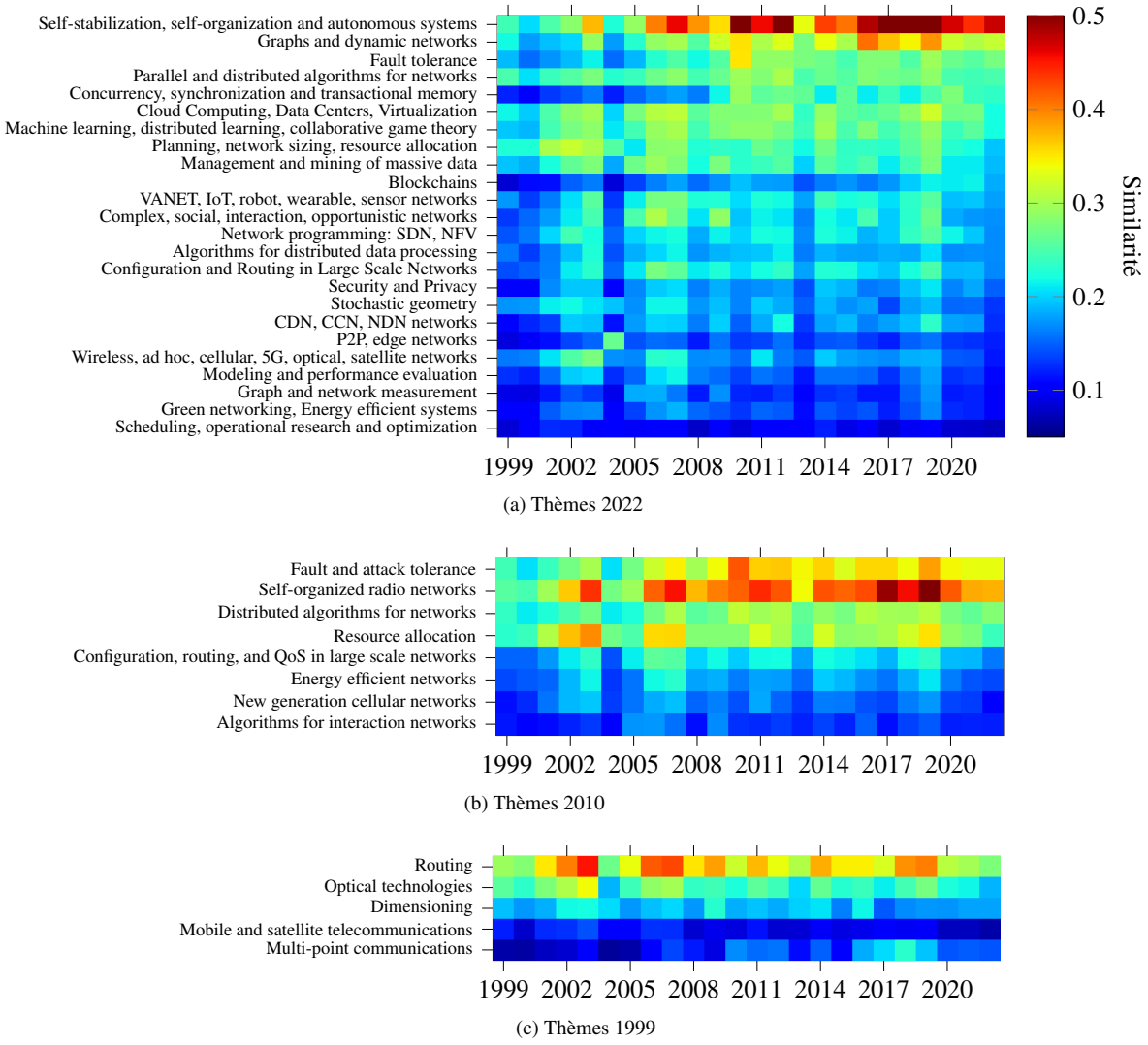
FIGURE 2 – Matrice des similarités entre les différentes éditions d'Algotel

s'expliquer par une forte réutilisation du PC de l'année précédente (au moins 45% dans chaque cas). On peut noter deux plus grosses zones de similarité (2000-2003, 2006-2022) qui correspondent d'une part à la période de création, et d'autre part à la période actuelle. Il est intéressant de remarquer que la période actuelle comprend également deux périodes de plus forte similarité (2009-2013, 2016-2022), malgré des taux de réutilisation du PC inférieurs à 20% (pour la période 2009-2013, et en 2016-2017) : une communauté scientifique peut donc rester stable même si ses représentants (le PC) évoluent fortement, si ceux-ci restent représentatifs de la communauté. L'année 2004 marque une rupture assez nette, qui peut s'expliquer par une forte volatilité du PC (un seul membre reconduit de l'année précédente, deux membres reconduits en 2005), qui entraîne une forte évolution de la communauté induite.

La figure 3 représente l'évolution des thématiques mises en avant dans l'appel à contributions d'Algotel vis-à-vis de la communauté scientifique Algotel depuis sa création. Nous avons considéré les thématiques mises en avant lors des éditions 2022, 2010, et 1999, traduites en anglais afin de les relier aux titres des articles publiés dans DBLP. Le tri des thématiques se fait par similarité décroissante vis-à-vis de l'année considérée. On peut faire les observations suivantes. Dans chacun des cas, environ la moité des thématiques mises en avant a une similarité faible avec la communauté scientifique Algotel, même si l'augmentation des thématiques au fur et à mesure des années permet de lisser ce phénomène. Certaines thématiques semblent même très peu en adéquation avec la communauté (par exemple « Mobile and satellite communication » en 1999, « Algorithms for interaction networks » en 2010, « Scheduling, Operational Research, and Optimization » en 2022). À l'inverse, on peut constater la bonne tenue dans le temps de thématiques historiques (« Routing » en 1999) et la montée d'autres thématiques (« Self-stabilization, self-organization and autonomous systems » en 2022). On observe à nouveau que la communauté de 2004 est faiblement similaire aux thématiques mises en avant dans les trois figures.

4 Conclusion

Nous avons montré comment, à partir de la seule connaissance des membres des comités de programme, des appels à contributions et de la base de donnée DBLP, il est possible d'analyser les thèmes et communautés d'une conférence et d'observer leur évolution. Au-delà de l'aspect analytique, notre approche peut également être utilisée pour assister à l'élaboration d'un comité de programme, en particulier pour éviter une faible similarité entre la communauté scientifique induite par le comité de programme et les thématiques mises en avant, comme observé Section 3. Le code source qui accompagne cet article (Section 2) comprend en particulier des outils pour obtenir des suggestions de membres du PC pertinents pour des thématiques (possiblement nouvelles), en fixant un taux de renouvellement.



(a) Thèmes 2022

(b) Thèmes 2010

(c) Thèmes 1999

FIGURE 3 – Adéquation entre thèmes annoncés et comité de programme.

Références

[1] Akiko Aizawa. An information-theoretic perspective of TF-IDF measures. *Info. Proc. & Manag.*, 39(1) :45–65, 2003.

[2] Marc-Olivier Buob and Fabien Mathieu. Gismo : Mettez un tigre dans votre moteur. In *Algotel*, 2020. `https://gismo.readthedocs.io/en/latest/`.

[3] Bruno Gaume and Fabien Mathieu. PageRank Induced Topology for Real-World Networks. working paper or preprint, May 2016.

[4] John H. Marburger III, Julia I. Lane, Stephanie S. Shipp, and Kaye Husbands Fealing, editors. *The Science of Science Policy : A Handbook*. Stanford University Press, 2011.

[5] Michael Kuhn and Roger Wattenhofer. The theoretic center of computer science. *SIGACT News*, 38(4) :54–63, 2007.

[6] Martin Rosvall and Carl T. Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4) :1118–1123, 2008.