Analyse géographique de séries de publications : application aux conférences EGC

Eric Kergosien, Marie-Noelle Bessagnet, Christian Sallaberry, Annig Le Parc-Lacayrelle, Albert Royer

Analyse géographique de séries de publications : application aux conférences EGC

Eric Kergosien*, Marie-Noëlle Bessagnet**,
Christian Sallaberry**, Annig Le Parc - Lacayrelle **, Albert Royer **

*GERiiCO, Université de Lille 3, 59653, Villeneuve d'Ascq Cedex
eric.kergosien@univ-lille3.fr
**LIUPPA, Université de Pau, 64013 Pau cedex
prénom.nom@univ-pau.fr

Résumé. Dans cet article, nous présentons une méthodologie originale permettant de faire des analyses scientométriques basées sur trois dimensions (spatiale, temporelle et thématique) à partir d'un corpus de publications. Cette méthodologie comporte 3 étapes : (1) la préparation et la validation des données pour compléter les critères usuels tels que les noms d'auteurs, affiliation, ... par des critères spatiaux, temporels et thématiques ; (2) l'indexation des contenus des publications et métadonnées associées ; (3) l'analyse et/ou la recherche d'information multidimentionnelle. Les expérimentations sont menées sur la série de publications des conférences EGC de 2004 à 2015.

1 Introduction

Nous assistons à un accroissement prodigieux des publications scientifiques disponibles au format numérique, que ce soit à l'échelle nationale ou internationale. Ce que promet la société du numérique est une toute autre façon de représenter et de concevoir l'espace et le temps : c'est notamment le cas des travaux en scientométrie.

Nos travaux s'inscrivent pleinement dans cette démarche en proposant une méthodologie semi-automatique pour l'analyse de l'évolution dans le temps et dans l'espace (1) d'un ensemble de publications scientifiques et (2) des thématiques concernées. L'intérêt de ces travaux est notamment d'appuyer les scientifiques dans leur travail de veille en mettant en avant l'évolution des thématiques au fil du temps, selon les lieux des conférences et les lieux des laboratoires d'affiliation des auteurs. Bien que de nombreux travaux en scientométrie présentent des méthodes pour analyser des communautés à partir de publications scientifiques (que ce soit en revues ou en conférences), il n'existe pas à notre connaissance de travaux proposant une analyse géographique d'un corpus de publications, i.e. combinant les dimensions spatiale, temporelle et thématique. Notre méthode pour l'analyse du corpus se décompose en 3 étapes : (1) préparation et validation du corpus pour le marquage géographique, (2) indexation du contenu des publications et des métadonnées associées, (3) analyse semi-automatique du corpus et recherche d'information (RI) multidimentionnelle. En fonction du corpus, le travail de préparation est plus ou moins conséquent et peut impliquer une action de saisie manuelle. Nous expérimenterons notre méthode sur le corpus de 1103 publications présentées à EGC

dans la période 2004-2015 en réponse au Défi EGC'2016. Nous proposons également un prototype de recherche d'information combinant des critères de recherche spatiale, temporelle, thématique et/ou plein texte.

Dans la deuxième partie de cet article, nous abordons les concepts théoriques importants relatifs à notre approche : la scientométrie et la recherche d'information géographique. Dans la troisième partie, nous présentons notre démarche générale applicable à différentes séries de publications scientifiques. Puis, la quatrième partie présente la mise en oeuvre de notre méthodologie sur les données des conférences EGC. Nous verrons qu'une étape de validation des méta-sessions basée sur une approche de fouille de textes a été mise en place. Enfin, nous terminerons par une conclusion et les perspectives de ces travaux.

2 Travaux connexes

La scientométrie se réfère à l'étude de tous les aspects de la littérature liée aux sciences et à la technologie (Hood et Wilson, 2001). Cela implique des analyses quantitatives sur les activités scientifiques, notamment les publications. Ainsi, on tente d'apprécier divers critères tels que l'évolution des pratiques des chercheurs, le rôle des sciences et de la technologie sur les économies nationales, l'évolution des technologies, etc. Il existe aujourd'hui des sources d'information publiques sur les publications scientifiques telles que Google Scholar, les archives ouvertes (HAL par exemple) permettant notamment d'analyser la production des chercheurs. Nous pouvons également accéder à des bases de données bibliographiques (SCOPUS, AERES, ...) pour enrichir les analyses. Dans le domaine de l'informatique, des analyses scientométriques ont été menées pour connaître l'évolution des pratiques des chercheurs concernant les articles publiés sur un corpus tel que la base de données DBLP (Cavero et al., 2014) ou encore évaluer les collaborations des chercheurs dans le cadre de leurs publications (Cabanac et al., 2015). Des travaux en France sont initiés dans le cadre d'un projet d'envergure dans le but d'acquérir et indexer des archives scientifiques pour créer une bibliothèque numérique. Ce projet dénommé ISTEX [1] a pour objectif de créer des services de recherche d'information innovants pour accéder à l'ensemble de ces ressources numériques selon différents critères. Dans sa version actuelle, l'application propose seulement d'accéder aux corpus par nom de revues ou par domaine scientifique.

La RI géograhique (RIG), nommée et définie pour la première fois par Ray Larson (Larson, 1996), est la tâche de recherche de documents satisfaisant des caractéristiques géographiques, la notion de géographie associant de façon explicite la dimension temporelle à la dimension spatiale et/ou thématique ((Martins et al., 2007) et (Liu et al., 2010)). À notre connaissance, de tels travaux de RIG n'ont jusqu'à présent pas été associés à ceux de scientométrie. Cependant, la combinaison ces trois dimensions semble importante pour améliorer les analyses.
Un récent travail intitulé « Anthologie des congrès Inforsid » [2] présente une analyse des 30 éditions du congrès. Il liste les thèmes de la conférence au fil des années et les villes des conférences et des laboratoires d'affiliation des auteurs. Enfin, une base de données permet d'accéder aux informations relatives aux articles et aux auteurs.

Au delà des analyses de ce premier travail, dédiées à la série de conférences INFORSID, nous proposons une approche générique applicable aux données relatives à toute nouvelle sé-

1. http ://www.istex.fr/le-projet/
2. *http ://dbrech.irit.fr/rechpub/cabanac_inforsid.accueil/*

rie de conférences. Notre approche supporte des opérations d'analyse et de recherche d'information combinant les dimensions spatiale, temporelle et thématique. Le modèle de données correspondant est illustré dans la section suivante.

3 Objectifs, démarche générale et mise en œuvre

Nous nous sommes fixés l'objectif d'observer, d'un point de vue spatial et temporel, des caractéristiques des auteurs et des thèmes des articles publiés dans une série de conférences.

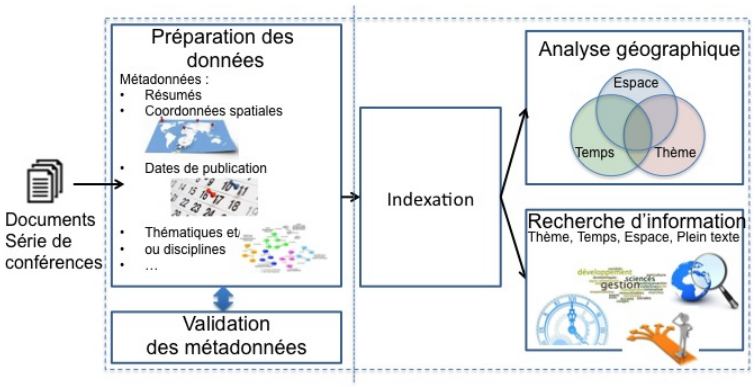

FIG. 1: Chaîne de traitement générique pour l'analyse de corpus de publications scientifiques.

Notre méthodologie est décrite par la figure 1. Indépendamment de tout corpus de publications scientifiques, la chaîne de traitement nécessite la présence de données spatiales, temporelles et thématiques. Dans le cas où ces données ne sont pas présentes, une étape de préparation est alors nécessaire pour les ajouter aux métadonnées existantes. Ainsi, pour chaque article, nous avons besoin :

— du point de vue spatial, du nom, de la latitude et la longitude des villes des auteurs et de la conférence ;
— du point de vue temporel, de l'année de sa publication ;
— du point de vue thématique, en plus de la liste des auteurs et du titre, le domaine traité ainsi que la session dans laquelle il a été présenté.

Nous validons les métadonnées spatiales, thématiques et temporelles en nous appuyant respectivement sur des outils de géocodage, de fouille de textes et une base calendaire.

Une deuxième étape concerne l'indexation de ces métadonnées dans un moteur de recherche afin de les exploiter dans des stratégies d'analyse et de recherche d'information combinant des critères spatiaux, temporels et thématiques. Ainsi, pour chaque article, l'index comprend les données listées dans le tableau 1. Il est à noter que l'ensemble des termes du titre et du résumé d'un article sont ajoutés à l'index dans la catégorie « Information plein texte » en vue de la RI.

Une troisième étape définit des processus génériques d'analyse et de présentation de données, applicables à des séries de conférences. L'analyse porte, de façon disjointe ou combinée,

Information thématique	Nom ville conférence
	Noms villes auteurs
	Noms auteurs
	Titre article
	Résumé article
	Session
	Domaine
Information spatiale	Coordonnées villes auteurs
	Coordonnées ville conférence
Information temporelle	Année conférence
Information plein texte	Termes titre article
	Termes résumé article

TAB. 1: Modèle synthétique des données indexées pour chaque article.

sur les dimensions spatiale, temporelle et thématique des données indexées à l'étape précédente. Par exemple, nous souhaitons observer les localisations des publiants par rapport aux localisations et aux années des conférences.

Une question, relative à la série de conférences observées, serait *y-a-t-il une modification de la distance moyenne des publiants par rapport à la localisation de la conférence ?* Ou encore, *y-a-t-il une forte proportion d'articles publiés par des co-auteurs distants géographiquement les uns des autres ?*

De plus, nous observerons les thématiques des conférences. *Peut-on exploiter le vocabulaire pour classer les articles à partir de leur résumé et définir des méta-sessions ? Quelle est la répartition des articles par méta-session ?*

Enfin, nous observerons également les conférences selon les dimensions spatiale, temporelle et thématique. *Quelle est la répartition, dans le temps et l'espace, des auteurs traitant de ces thématiques ? Y-a-t-il des spécificités thématique par région, par période ?*

Ces processus s'appuient principalement sur le moteur de recherche ElasticSearch [3] mais également sur des outils tels que MapInfo [4] et Google Earth. ElasticSearch est un moteur de recherche basé sur la librairie Lucene. Il permet la recherche plein-texte, la recherche structurée, l'analyse des documents qu'il a indexés et la gestion des données spatiales.

4 Expérimentations sur les données de EGC

Nous avons appliqué ces processus et mis en œuvre ces outils pour l'observation de la série de conférences EGC.

4.1 Préparation et validation des données

Dans le cadre du défi EGC, les données disponibles dans le fichier « RNTI_articles_export.txt » (fourni par le Défi) concernent les articles publiés à la conférence EGC depuis 2004. Nous appellerons ce fichier « articles » dans la suite de cet article. Pour chacun des articles publiés dans ces conférences, ce fichier fournit les informations suivantes :

3. https ://www.elastic.co/fr/
4. http ://www.pitneybowes.fr/software/geo-decisionnel-sig/mapinfo-suite/mapinfo-professional.shtml

— year : année de publication (temporel),
— title : titre de l'article ; abstract : résumé de l'article ; author : noms d'auteur séparés par une virgule (thématique),
— series : nom du journal ; booktitle : nom de la conférence ; pdf1page : lien hypertexte du fichier pdf pour la première page ; pdfarticle : lien hypertexte du fichier pdf pour l'article.

Un premier traitement a permis de filtrer les données du fichier afin d'en exclure toutes les lignes des manifestations non EGC. Les traitements suivants visent la préparation des données spatiale et thématique. Ici, aucun traitement n'a été nécessaire pour la préparation des données temporelles, déjà renseignées dans la version initiale du fichier « articles ».

4.1.1 Dimensions spatiale et temporelle

Une première étape consiste à compléter les informations relatives à chaque article, en ajoutant automatiquement le lieu de la conférence et manuellement les noms des villes des auteurs (affiliation du laboratoire). La seconde étape, quant à elle, complète ces données par les coordonnées spatiales relatives aux différentes villes. Des services [5] [6] de géocodage nous ont permis de déterminer, automatiquement, les latitudes et longitudes des villes d'affiliation des auteurs et des conférences.

4.1.2 Dimension thématique

La composante thématique correspond aux principaux domaines traités dans le cadre de la conférence EGC, domaines explicités dans le nom des sessions proposées entre 2004 et 2015. Pour traiter cette dimension, les métadonnées mises à disposition ne sont pas suffisantes. En effet, les sessions durant lesquelles les articles sont présentés ne figurent pas dans le fichier «articles». Nous avons enrichi manuellement le jeu de données en renseignant les sessions pour chaque article. Les sessions traitant un même domaine ne portant pas le même nom au fil des années, nous avons analysé l'ensemble des sessions pour les regrouper par domaine d'études que nous nommons méta-sessions. Un ensemble de 8 méta-sessions a ainsi été défini (cf. tableau 2). Par exemple, les sessions « Gestion des connaissances » (2004), « Construction et alignement d'ontologies » (2011) ou encore « Sémantique et Ontologies » (2015) ont été affectées à la méta-session 1 « Organisation des connaissances ». Ainsi, pour considérer cette dimension, nous avons tout d'abord classé les sessions dans les 8 méta-sessions, puis nous avons affecté les articles aux méta-sessions correspondantes.

Sur les 1103 articles recensés de 2004 à 2015, nous ne pouvons prendre en compte les posters et les conférences invitées car ils ne sont pas intégrés dans une session. Nous ne traitons pas les articles en anglais car notre approche de fouille de textes pour validation est sensible à la langue. Le corpus analysé dans nos travaux est donc constitué de 812 articles, soit environ 73% du corpus initial.

Validation des méta-sessions Dans le but de valider le regroupement des sessions en méta-sessions, nous proposons une méthode incrémentale de fouille de textes s'appuyant sur les

5. *https ://adresse.data.gouv.fr/tools/*
6. *http ://www.batchgeocodeur.mapjmz.com/*

Méta-session	Identifiant	Nombre articles
Organisation des connaissances	1	129
Apprentissage	2	69
Séquences, Motifs et règles d'association	3	170
Classification	4	82
Logiciels et applications	5	122
Visualisation et RI	6	62
Web et réseaux sociaux	7	26
Données et Big data	8	152

TAB. 2: Méta-sessions définies pour les publications des conférences EGC entre 2004 et 2015.

résumés de chaque article (cf. le champ abstract) pour classifier automatiquement l'ensemble des articles dans les méta-sessions. Nous faisons l'hypothèse que le vocabulaire utilisé dans les résumés est suffisamment discriminant pour identifier la méta-session correspondant à un article. La catégorisation de textes se décompose en trois grandes familles, à savoir la proposition d'algorithmes ad hoc (Turney, 2002) (Snyder et Barzilay, 2007), l'utilisation de ressources lexicales (Kim et Hovy, 2004), ou encore les méthodes d'apprentissage telle que Naive Bayes (Su et al., 2008). Les méthodes d'apprentissage supervisé peuvent être plus ou moins complexes : (i) tous les mots du texte (sac de mots, unigrammes ou ngrammes) (Pang et al., 2002) ; (ii) la présence ou l'absence d'un ensemble de mots déterminés ; (iii) l'emplacement de certains mots (Kim et Hovy, 2006). Au regard du corpus annoté intégrant les informations sur les sessions et méta-sessions, nous nous positionnons dans une démarche de classification supervisée pour (in-)valider de façon incrémentale les méta-sessions définies manuellement. Les résultats sont analysés en utilisant la mesure de performance F-mesure. Aussi, afin d'estimer l'efficacité des différentes étapes, nous appliquons un processus de validation croisée, méthode d'estimation de fiabilité d'un modèle fondée sur une technique d'échantillonnage. Le tableau 3 présente les résultats obtenus à partir de la 1re série de tests, en terme d'exactitude (accuracy). La méthode mise en place se décompose en 3 étapes :

1. approche sac de mots classique en utilisant l'algorithme de fouille de données Naives Bayes DMNBtext (Su et al., 2008) : l'ensemble des mots d'un résumé forme un vecteur signature qui est utilisé pour identifier à quelle méta-session doit être attaché l'article correspondant ;
2. nettoyage du corpus :
   (a) lemmatisation : Prise en compte des lemmes uniquement,
   (b) suppression des mots vides,
3. pondération des mots : pour chaque méta-session, création d'une liste de mots constituant le nom des sessions attachées. Nous attribuons le poids de 1 aux mots lemmatisés du vecteur signature d'un article non présent dans ces listes et 10 à ceux présents (valeur la plus pertinente au vu de l'ensemble des expérimentations réalisées).

Les expérimentations ont tout d'abord été réalisées sur l'ensemble du corpus et les résultats obtenus en nous appuyant sur l'approche sac de mots classique (cf. Tableau 3, étapes 1 et 2) sont mitigés. Nous expliquons ces résultats par le fait que le vocabulaire a évolué entre 2004 et 2015 et que ce vocabulaire dans son ensemble n'est pas discriminant. Lorsqu'on pondère

les mots en prenant en compte le fait qu'ils appartiennent aux noms des sessions, les résultats sont meilleurs avec un score de 84.8% de classification correcte avec une validation croisée de 15 plis (cf. Tableau 3, étape 3). Aussi, le nombre de publications diffère au fil du temps pour chaque méta-session et certaines méta-sessions sont peu ou parfois pas traitées sur l'ensemble des conférences EGC entre 2004 et 2015. C'est notamment le cas de la méta-session « Web et réseaux sociaux ». Nous proposons de décomposer notre corpus en sous-ensembles correspondants à des périodes plus restreintes dans le temps, mais homogènes sur le nombre d'articles : un premier sous-ensemble constitué des articles publiés entre 2004 et 2007, un second entre 2008 et 2011 et le troisième entre 2012 et 2015. Nous avons relancé les expérimentations sur chacun de ces sous-ensembles et les résultats sont nettement meilleurs, avec des taux de classification valides de 81,9% à 95,7% (cf. Tableau 3, étape 4). Nous remarquons que les meilleurs résultats sont obtenus avec une validation croisée de 10 plis.

Etape	Description	Score F-mesure (Validation croisée avec N plis)			
		N=3	N=5	N=10	N=15
1	Approche sac de mots classique	0.197	0.277	0.301	0.321
2*2	a. Lemmatisation	0.38	0.396	0.394	0.394
	b. Suppression des mots vides	0.244	0.393	0.387	0.387
3	Pondération des mots	0.741	0.776	0.822	**0.848**
3*4	Pondération des mots pour la période 1	0.649	0.756	**0.819**	0.814
	Pondération des mots pour la période 2	0.812	0.891	0.914	**0.920**
	Pondération des mots pour la période 3	0.793	0.862	**0.957**	0.951

TAB. 3: Détails des résultats pour la validation des méta-sessions.

Ces résultats nous permettent de valider la classification des articles de la conférence en 8 méta-sessions. Nous pouvons analyser l'évolution des publications par thème dans le temps et dans l'espace.

4.2 Indexation

À l'issue de la phase de préparation nous disposons du fichier « articles_préparés ». Pour chaque article décrit dans ce fichier, un document json, conforme au modèle exposé tableau 1, est généré. Ce document est ensuite indexé sous ElasticSearch.

4.3 Analyse des données

Nous avons mené un grand nombre d'analyses [7] dont nous présentons quelques résultats.

4.3.1 Dimensions spatiale et temporelle

Grâce à des services [8] [9] de géocodage et fonctions de calcul de distance, nous analysons les données relatives à la dimension spatiale (calcul des distances moyennes entre les villes

7. http ://ekergosien.net/DefiEGC
8. *https ://adresse.data.gouv.fr/tools/*
9. *http ://www.batchgeocodeur.mapjmz.com/*

des auteurs d'un article et celle de la conférence correspondante, ou encore, entre les villes des auteurs d'un article donné).

Nous avons observé la localisation des publiants par rapport au lieu et à l'année de la conférence. Nous présentons la moyenne des distances entre les publiants de chaque article et la localisation de la conférence sur une cible (figure 2).

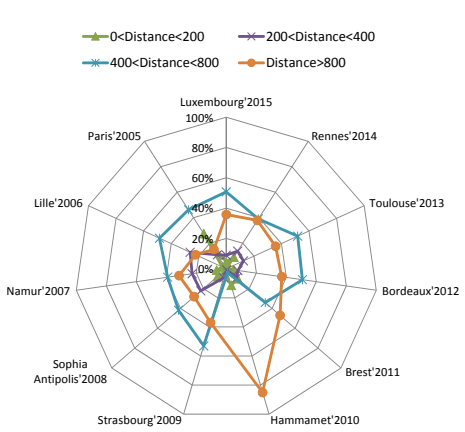

FIG. 2: Articles (en pourcentage) par distance moyenne (en km) du lieu de la conférence par rapport aux lieux d'affiliation des co-auteurs.

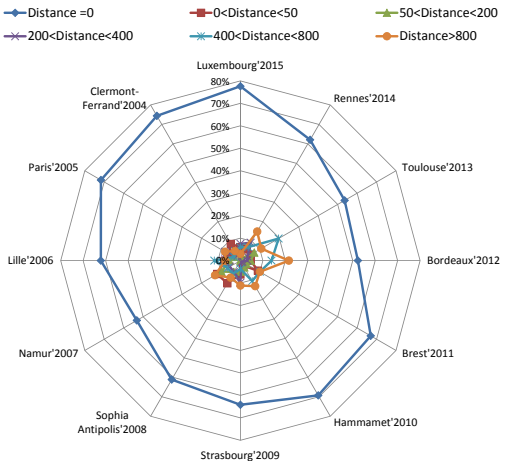

FIG. 3: Articles (en pourcentage) par distance moyenne (en km) des lieux d'affiliation des co-auteurs.

Ainsi, la série $200 < Distance < 400\ km$, montre une diminution considérable du pourcentage d'articles en 2010 et au-delà. La tendance est la même pour la série $400 < Distance < 800\ km$. D'autre part, la conférence d'Hammamet présente un pourcentage d'articles, exceptionnellement important, pour lesquels la distance moyenne des co-auteurs par rapport à la conférence est supérieure à $800\ km$. Enfin, la conférence de Paris présente un pourcentage d'articles, très important, pour lesquels la distance moyenne des co-auteurs par rapport à la conférence est inférieure à $200\ km$. Le constat est le même pour une distance moyenne comprise entre 400 et $800\ km$.

Nous avons également observé la distance moyenne entre les villes des co-auteurs de chaque article. Nous présentons la proportion d'articles publiés par tranches de distance géographique entre co-auteurs sur une cible (figure 3). La série $Distance = 0\ km$ montre que les co-auteurs sont majoritairement de la même ville quelle que soit la ville ou l'année de la conférence. Cette tendance est toutefois particulièrement accentuée pour les conférences de Paris, de Clermont-Ferrand et de Luxembourg.

4.3.2 Dimension thématique

Sur la figure 4, nous montrons l'évolution du classement des articles par méta-session sur les 12 années de conférence. Au fil des ans, seules deux méta-sessions perdurent : 1 (Organisation des connaissances) et 3 (Séquences, Motifs et règles d'association). La méta-session 8

(Données et Big data) est présente de 2004 à 2014 et disparaît en 2015. Ces trois méta-sessions se révèlent être les méta-sessions où l'on a le plus de publiants en cumul. La méta-session 6 (Visualisation et RI) est absente une seule année : 2009. D'ailleurs, en 2009, il ne reste plus que 5 méta-sessions. La méta-session 7 (Web et réseaux sociaux) est présente en 2004 et ne réapparaît qu'en 2010. C'est une méta-session où l'on publie peu (26 articles entre 2004 et 2015). Les méta-sessions 2 (Apprentissage) et 4 (Classification) ne sont pas présentes tous les ans : 8 années pour la première et 10 années pour la seconde. Quant à la méta-session 5 (Logiciels et applications), elle n'a été absente qu'une seule année (en 2008). C'est une méta-session possédant également un nombre de publiants significatif.

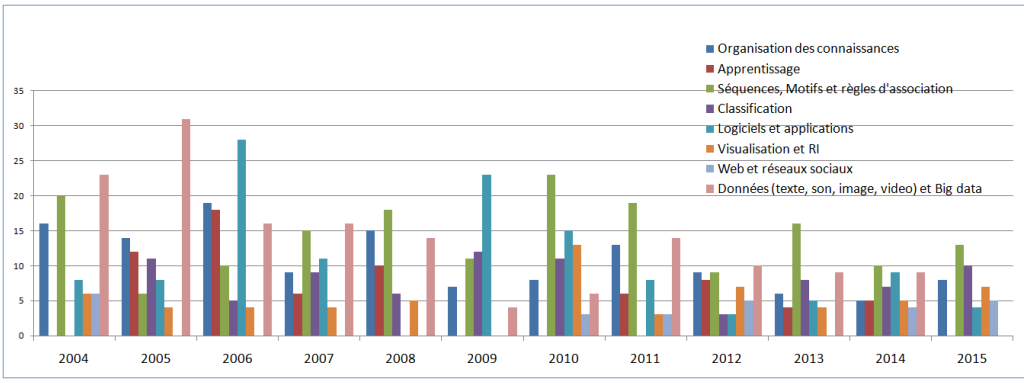

FIG. 4: Evolution du nombre d'articles publiés par méta-session dans le temps.

4.3.3 Prise en compte des trois dimensions

Nous avons observé la répartition géographique des articles pour chacune des méta-sessions sur 3 périodes distinctes : 2004-2007, 2008-2011 et 2012-2015 (cf. Figure 5). Pour cette analyse, nous avons gardé les villes les plus représentatives, i.e. nombre de publiants > 20 sur les 12 années de publication pour les 8 meta-sessions.

Au regard de la Figure 5, nous faisons les observations suivantes :

— la région parisienne, Lyon, Rennes et Montpellier fournissent le plus grand nombre de publiants sur les trois périodes,
— certaines régions (Lille, Grenoble) n'ont quasiment pas de publiants sur une période,
— certaines régions (Lille, Lyon, Nancy, Nice) n'ont pas de publiants pour une ou plusieurs métasessions,
— de nombreuses régions (par exemple, Toulouse, Nice, Grenoble, Nancy, Nantes) sont en perte de vitesse entre la première période et la dernière.

4.4 Recherche d'information géographique

Nous avons mis en place une méthode de recherche d'articles combinant la recherche plein texte (RI classique) avec la RI géographique. Nous avons testé différents scenarii (tableau 4). Seuls les articles ayant un résumé sont considérés. La première colonne du tableau décrit les

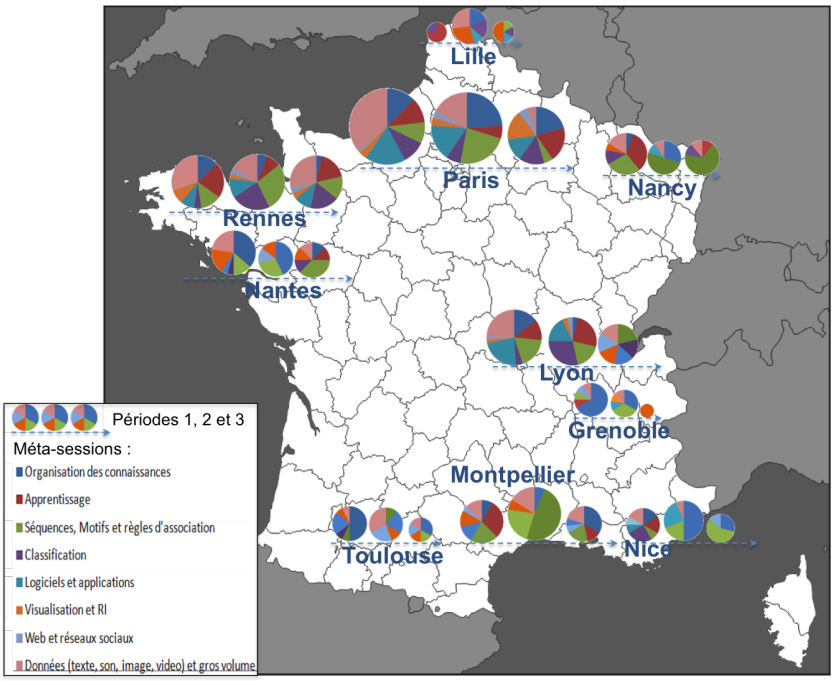

FIG. 5: Analyse géographique (temps - espace - thème) des publications.

différents scenarii pour la recherche plein texte (présence ou non du terme « classification » dans le titre et/ou le résumé). La deuxième colonne donne le nombre d'articles correspondant à chacun d'eux. Les colonnes suivantes correspondent à l'application de critères supplémentaires aux résultats de la recherche plein texte :

— le critère thématique considère les articles associés à la méta-session "Classification" ;
— le critère temporel considère les articles des conférences EGC présentés depuis 2007 ;
— le critère spatial considère les articles dont au moins un des auteurs est localisé dans une ville à moins de 200 km de Paris.

Recherche du terme "classification"		Critères appliqués			
		Thème	Thème Temps	Thème Espace	Thème Espace Temps
$titre \wedge \neg résumé$	17	5	5	0	0
$\neg titre \wedge résumé$	100	21	17	9	6
$titre \wedge résumé$	79	23	19	11	9
$titre \vee résumé$	**196**	**49**	41	20	15
$\neg titre \wedge \neg résumé$	752	30	25	12	11
aucune recherche plein-texte	948	**79**	66	32	26

TAB. 4: Nombre d'articles répondant aux différents scénari de recherche

Nous constatons (tableau 4) que 79 articles relèvent de la méta-session « Classification », que 196 contiennent le terme « classification » dans le titre ou le résumé et que 49 sont à la fois dans la méta-session « Classification » et contiennent ce terme dans le titre ou le résumé. Nous pouvons en conclure que la recherche plein-texte et la recherche thématique par méta-session sont complémentaires dans ce cas précis (comme dans d'autres cas observés). Nous constatons également que les critères spatial et temporel apportent un niveau de précision supplémentaire dans l'observation des données.
Dans cette expérimentation, nous avons donné le même poids à chaque critère de recherche et les articles retournés sont classés par ordre de pertinence. Notons toutefois qu'il est possible de pondérer ces critères pour privilégier ou bien défavoriser un axe de recherche donné.

5 Conclusion et perspectives

Dans cet article, nous avons présenté une démarche d'analyse de publications scientifiques considérant quatre dimensions différentes : une dimension plein texte, une dimension thématique, une dimension spatiale et une dimension temporelle. Les résultats présentés, appliqués aux publications des conférences annuelles EGC de 2004 à 2015, montrent que grâce à la combinaison de ces 4 dimensions, nous faisons une analyse plus fine et moins classique que dans les travaux menés jusqu'alors en scientométrie. En effet, concernant la dimension thématique, nous avons classé les divers articles selon des méta-sessions (regroupement de sessions par thèmes), classification que nous avons validée par une approche de fouille de textes en nous appuyant sur les résumés des publications. Les dimensions spatiale et temporelle ont permis de mettre en place des filtres montrant des points de vue originaux sur les analyses statistiques. La combinaison de ces dimensions nous permet de proposer un point de vue géographique du corpus traité. Notre méthodologie d'analyse et de recherche d'information est générique. En effet, les étapes de préparation de données, d'indexation, d'analyse et de recherche d'information, pourront être appliquées quelques soient la série de conférences et les données associées. Les perspectives sont variées. D'une part, lors de l'étape de préparation de données spatiales, temporelles et thématiques, nous avons intégré des données supplémentaires nécessaires à nos analyses. Les traitements manuels pourront être automatisés, notamment en ce qui concerne la recherche des villes d'affiliation des laboratoires des auteurs. D'autre part, nous pourrons appliquer la démarche de classification sur les articles retenus à EGC 2016 pour (1) aider au classement de ces derniers dans les méta-sessions, et (2) les intégrer dans notre analyse géographique. À moyen terme, nous souhaitons appliquer notre méthodologie sur un autre corpus en langue anglaise pour tester et valider sa généricité.

Références

Cabanac, G., G. Hubert, et B. Milard (2015). Academic careers in computer science : continuance and transience of lifetime co-authorships. *Scientometrics 102*(1), 135–150.

Cavero, J., B. Vela, et P. Cáceres (2014). Computer science research: more production, less productivity. *Scientometrics 98*(3), 2103–2111.

Hood, W. W. et C. S. Wilson (2001). The literature of bibliometrics, scientometrics, and informetrics. *Scientometrics 52*(2), 291–314.

Kim, S.-M. et E. Hovy (2004). Determining the sentiment of opinions. In *Proceedings of the 20th International Conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kim, S.-M. et E. Hovy (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, Stroudsburg, PA, USA, pp. 1–8. Association for Computational Linguistics.

Larson, R. R. (1996). Geographic Information Retrieval and Spatial Browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, 81–124.

Liu, X., C. Jian, et C.-T. Lu (2010). A spatio-temporal-textual crime search engine. In *GIS'10 : Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '10, New York, NY, USA, pp. 528–529. ACM.

Martins, B., J. Borbinha, G. Pedrosa, J. a. Gil, et N. Freire (2007). Geographically-aware information retrieval for collections of digitized historical maps. In *GIR'07: Proceedings of the 4th ACM workshop on Geographical information retrieval*, New York, NY, USA, pp. 39–42. ACM.

Pang, B., L. Lee, et S. Vaithyanathan (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, Stroudsburg, PA, USA, pp. 79–86. Association for Computational Linguistics.

Snyder, B. et R. Barzilay (2007). Multiple aspect ranking using the good grief algorithm. In *In Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL*, pp. 300–307.

Su, J., H. Zhang, C. X. Ling, et S. Matwin (2008). Discriminative parameter learning for bayesian networks. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, New York, NY, USA, pp. 1016–1023. ACM.

Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Stroudsburg, PA, USA, pp. 417–424. Association for Computational Linguistics.

Summary

In this paper, we present an original methodology making scientometric analysis based on three dimensions (spatial, temporal and thematic). This methodology includes 3 steps: (1) preparation and validation of data in order to complete usual criteria such as author names, affiliation, ... by spatial, temporal and thematic ones; (2) the indexating of the contents of the publications and associated metadata; (3) analysis and/or multidimensional information retrieval. Expriments are conducted on EGC conferences from 2004 to 2015.