Mots-clefs, Collaborations, Sentiments: Évolutions de la Conférence EGC depuis 2004

Erick Stattner, Didier Henry, Nathan Jadoul

Université des Antilles
Laboratoire de Mathématiques, Informatique et Applications (LAMIA)
{erick.stattner, didier.henry, nathan.jadoul}@univ-antilles.fr

Résumé. Dans cet article, nous menons un travail d'analyse de données issues de la conférence "*Extraction et Gestion des Connaissances*" (EGC). Ces données concernent les articles acceptés à la conférence EGC depuis 2004 ainsi que les messages publiés sur le réseau social Twitter. Notre objectif est de mettre en évidence, depuis 2004, les grandes évolutions de la conférence EGC en matière (i) de mot-clefs, (ii) de collaborations et (iii) de sentiments exprimés sur Twitter. Les résultats obtenus ont ainsi permis de mettre en lumière l'évolution des thématiques de la conférence depuis 2004, les schémas collaboratifs qui se mettent en place ainsi que leur évolution, et les sentiments qui émergent des messages publiés autour de la conférence sur Twitter.

1 Introduction

L'extraction de connaissances de données hétérogènes est devenue un des axes de recherche les plus prolifiques de ces dernières décennies (Waller et Fawcett, 2013; Van Der Aalst, 2016). Cette explosion des travaux menés autour de l'analyse de données s'explique par plusieurs facteurs. (i) Tout d'abord la collecte de données est aujourd'hui massive. En effet, des données très hétérogènes peuvent aujourd'hui être collectées : données de terrain, données issues des réseaux sociaux en ligne, données capturées par des périphériques mobiles, données individuelles induites ou calculées à partir d'autres données et même des données de projection simulées par des algorithmes. (ii) L'amélioration des méthodes d'analyse est également un facteur qui explique ce succès, puisque de nombreuses méthodes ont été proposées pour tenir compte de cette hétérogénéité des données et rechercher des schémas de connaissances de plus en plus adaptés au contexte. (iii) Enfin, la diversité des problèmes abordés et la précision des résultats sont sans doute l'un des aspects les plus marquants du domaine. En effet, des problèmes très variés peuvent aujourd'hui être abordés par l'analyse des données : impact de phénomènes de diffusion (Barrat et al., 2014), prédiction d'élections (Chung et Mustafaraj, 2011), détection d'évènements (Sakaki et al., 2010), etc.

Dans ce travail, nous menons un travail d'analyse de données issues de la conférence "*Extraction et Gestion des Connaissances*" (EGC). Ces données concernent les articles acceptés à la conférence EGC depuis 2004 ainsi que les messages publiés sur le réseau social Twitter. Notre objectif est de mettre en évidence, depuis 2004, les grandes évolutions de la conférence

EGC en matière (i) de mot-clefs (ii) de collaborations et (iii) de sentiments exprimés sur Twitter.

Pour ce faire, nous commençons par analyser chaque année les titres des articles scientifiques publiés à la conférence pour en extraire les mots-clés associés. Nous caractérisons ensuite les transitions qui s'opèrent sur les thématiques pour montrer leur évolution au cours du temps. Dans un second temps, nous adoptons une approche orientée réseau pour étudier les collaborations qui prennent place à la conférence et en particulier l'évolution des schémas collaboratifs qui semblent émerger dans le processus de collaboration. Enfin, nous analysons les données Twitter pour étudier l'évolution des opinions autour des messages postés sur l'évènement afin de montrer l'évolution des sentiments positifs et négatifs qui s'y dégagent.

L'article est structuré de la façon suivante. La section 2 décrit la méthode mise en place pour extraire les mot-clefs, ainsi que les résultats obtenus. La Section 3 est consacrée aux collaborations et détaille les schémas collaboratifs qui émergent des papiers soumis à la conférence. La section 4 est, elle, dédiée à l'étude des opinions sur Twitter. Enfin, la section 5 conclut et décrit les pistes futures.

2 Évolution des mot-clefs

Dans une première étape, nous nous sommes intéressés à l'évolution des mots-clefs. Pour ce faire nous avons étudié le corpus composé des titres des articles fournis par la conférence. Nous nous sommes en particulier focalisés sur les titres des articles. En effet, le titre d'un article de par sa fonction nécessite un travail de synthèse de l'information. Nous présenterons tout d'abord le modèle d'analyse de texte développé puis la phase de récolte et de nettoyage du corpus, et enfin les principes de notre analyse de texte ainsi que les résultats obtenus. Pour étudier l'évolution des mots-clefs des titres, nous nous concentrons sur les mots les plus fréquents de ceux-ci. Les dix mots les plus fréquents pour chaque année sont extraits.

Notre modèle permet l'étude de l'évolution de mots à travers plusieurs comportements. Ce modèle, consiste à analyser l'évolution de 4 comportements chez les mots qui lui sont présentés. Ces 4 comportements sont :

— la naissance de mot (quand un mot apparaît pour la première fois)
— la mort d'un mot (quand un mot disparaît définitivement)
— la réapparition d'un mot (quand un mot réapparaît après une période de disparition)
— la continuité d'un mot (quand un mot continu d'exister d'une année à l'autre)

Afin d'analyser les titres des articles, une phase de collecte et d'extraction a du préalablement être mise en place. Pour ce faire nous nous sommes aidé du fichier csv fourni par la conférence listant tous les articles parus de 2004 à 2018. Les titres ont directement été obtenus de ce fichier csv. Nous obtenons alors un sujet d'étude, des corpus composés des titres des articles triés par année. Le sujet est donc composé de 14 corpus, chacun représentant une année.

Pour le travail de prétraitement du texte, le langage R a été choisi en raison de ces nombreuses bibliothèques de text mining, pour cela, nous procédons comme suie : tout d'abord, nous avons supprimé la ponctuation et les chiffres du texte. Puis, nous avons mis le tout en minuscules et supprimé ce qu'on appelle les mots vides à l'aide d'une liste déjà implémentée dans R. Ensuite, nous ne gardons que la racine des mots restants (Ex : temps, temporel qui deviennent juste temp).

Maintenant que le texte a été nettoyé, la matrice des termes des documents (DTM) peut être générée. Elle nous permet de connaître le nombre d'occurrences d'un mot dans un texte. À partir de cette matrice, nous pouvons générer les 10 mots-clefs dans le corpus, pour cela, on somme la DTM sur ces colonnes et on obtient un vecteur qu'il suffit d'ordonner pour récupérer les termes les plus fréquents.

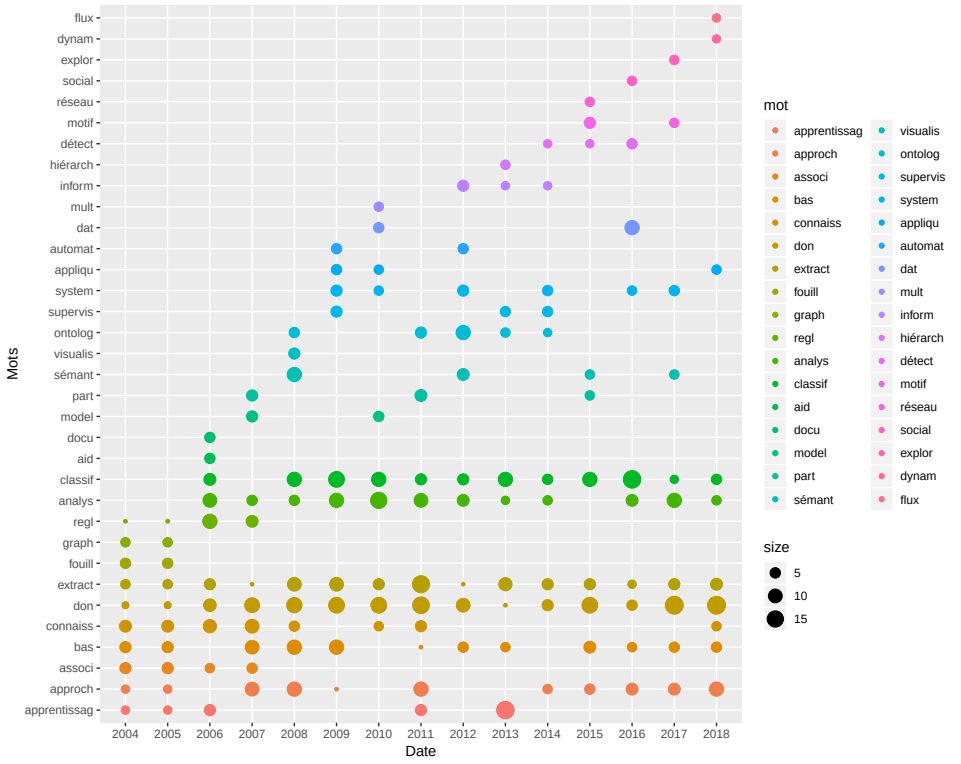


FIG. 1 – *Évolution globale des thèmes des titres*

L'évolution globale des mot-clés est illustrée par la Figure 1. Ce graphique présente par années les différents mot-clés identifiés et met en évidence le nombre d'occurrences du mot par la taille du point dans le graphique. Nous remarquons que les mots *extract* et *don* sont présents tout au long de l'étude et constitue alors les mots réguliers dans les titres des articles. Certains mots comme *approch*, *bas*, *analys* et *classif* se maintienne sur la majorité de l'étude et constitue aussi des mots réguliers. À partir des années 2010 apparaissent des mots qui ne sont propres qu'à leur année d'études et viennent diversifier les mots fréquents.

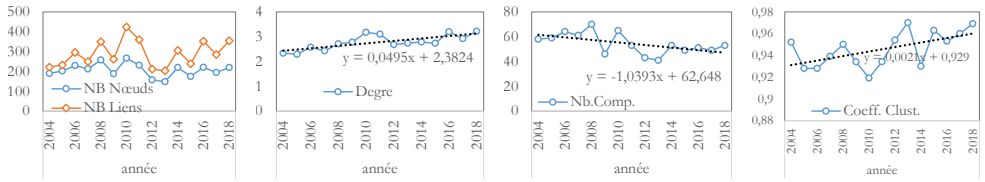
Puis nous nous sommes intéressés à certains comportement de notre modèle d'analyse. Nous avons constaté une période de diminution des naissances à partir des années 2010, avec en 2006 4 naissances et en 2012, une seule. Ensuite, il y a une période de relative stagnation en 2012 avec des naissances comprisent entre 1 et 2 par années. Concernant les morts des mot-clefs, nous avons remarqué une augmentation de mort à partir de 2013 avec 2 morts en 2006 et 3 morts en 2013 puis 4 morts en 2017.

3 Schémas collaboratifs

Dans cette section, nous nous intéressons aux collaborations et tentons d'identifier les schémas collaboratifs cachés au sein des articles acceptés à la conférence EGC.

Dans cet objectif, nous avons tout d'abord commencé par générer, pour chaque année, le "*réseau des collaborations*". Dans ce réseau, chaque noeud correspond à un auteur. Un lien existe entre deux auteurs, s'ils ont co-signé un papier. Ainsi, tous les auteurs d'un même papier sont liés entre eux et forment donc un sous-réseau complet au sein du réseau de collaborations. On pourrait ainsi penser que le réseau de collaborations n'est composé que de sous-réseaux non-connectés entre eux. Il est cependant important de prendre en compte qu'un auteur peut soumettre plusieurs papiers à la conférence avec des auteurs différents, ce qui crée des ponts entre les composantes du réseau et renforce ainsi sa densité.

La Figure 2 montre l'évolution des principales caractéristiques du réseau de collaborations sous-jacent : (a) taille du réseau (nombre de noeuds et liens), (b) degré moyen des auteurs, (c) nombre de composantes connexes et (d) coefficient de clustering.



(a) (b) (c) (d)

FIG. 2 – *Évolution des principales caractéristiques du réseau de collaborations (a) taille, (b) degré moyen, (c) nombre de composantes connexes et (d) coefficient de clustering*

Nous pouvons tout d'abord observer que la taille du réseau de collaborations a cru chaque année. En effet, le nombre d'auteurs est passé de 190 en 2004 à 220 en 2018. Le nombre de collaborations est, lui, passé de 222 à 354 sur la même période. L'évolution du degré moyen est particulièrement intéressante, puisqu'elle montre qu'en moyenne les auteurs ont accru leur nombre de collaborations, passant de 2,3 collaborations en moyenne par auteur en 2004 à 3,2 en 2018. Cet accroissement du nombre de collaborations s'observe également sur le nombre de composantes connexes en diminution. En effet, si nous pouvions nous attendre à un taux élevé de composantes dans un tel réseau, il est important de noter qu'environ 1 composante est perdue chaque année. Ce qui suggère que certains auteurs soumettent plusieurs papiers avec des auteurs différents, créant ainsi des ponts entre des auteurs de différentes composantes et renforçant l'effet communautaire du réseau. Cet effet communautaire peut également s'observer sur l'augmentation du coefficient de clustering moyen du réseau.

Dans un second temps, nous nous sommes intéressés aux schémas collaboratifs cachés dans ces réseaux de collaborations. Pour ce faire, nous avons commencé par extraire pour chaque auteur 9 attributs. (1) identifiant, (2) nb. papiers, (3) nb. total collaborations, (4) nb. moyen collaborations par papiers, (5) position moyenne dans liste des auteurs, (6) nombre de fois ou il est premier auteur, (7) terme le plus présent dans ses articles, (8) second terme le plus présent, et (9) troisième terme le plus présent.

Nous avons ensuite extrait du réseau de collaborations de chaque année, les liens conceptuels (Stattner et Collard, 2012), pour mettre en évidence les liens fréquents entre les groupes de noeuds. Cette technique, qui effectue du clustering de liens, recherche dans les liens du réseau les schémas collaboratifs, c'est-à-dire les groupes de noeuds (définis par un ensemble d'attributs) fréquemment connectés à d'autres groupes de noeuds (également définis par un ensemble d'attributs). La Figure 3 montre les réseaux de collaborations en (a) 2004, (b) 2011 et (c) 2018, ainsi que les schémas de collaborations associés (d), (e) et (f) obtenus avec un seuil de support de 20%

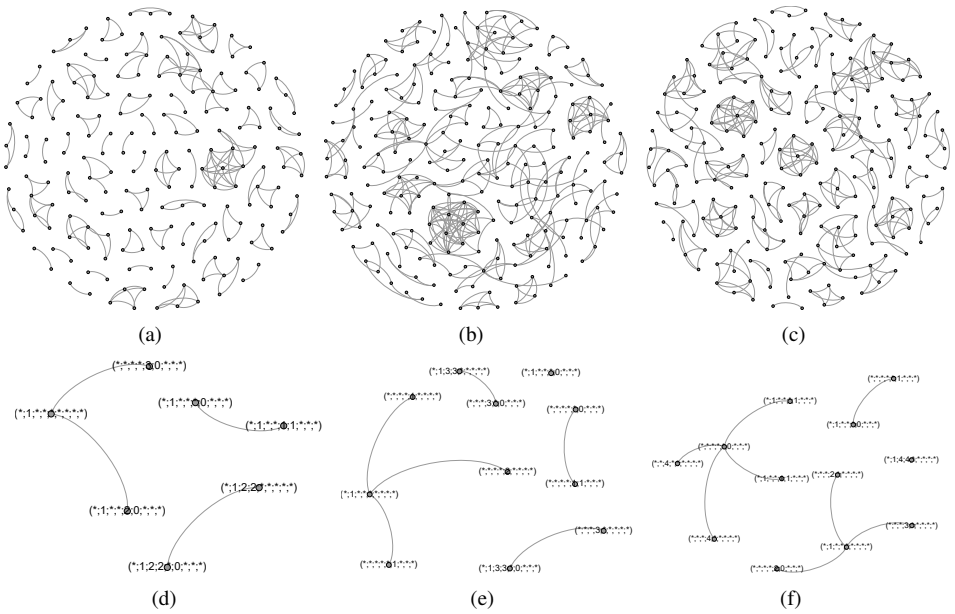


FIG. 3 – *Réseaux des collaborations en (a) 2004, (b) 2011 et (c) 2018 et clusters de collaborations associés (d), (e) et (f) obtenus avec un seuil de support de* 20%

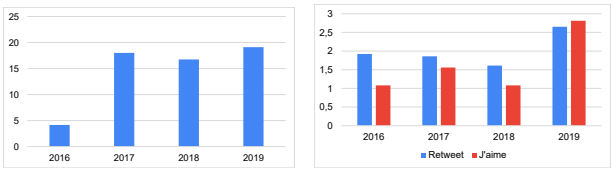
Comme observé précédemment sur l'évolution du nombre de composantes connexes (cf Figure 2(c)), on peut noter sur les figures 3(a), (b) et (c) la densification des liens du réseau de collaborations dans le temps. Concernant les schémas collaboratifs, il est intéressant d'observer, pour chaque année, que des schémas peuvent être extraits avec un seuil de 20%, ce qui montre que des motifs forts peuvent être trouvés des les collaborations des articles publiés à EGC. Par exemple, en 2004, nous obtenons le schéma suivant :

$$\mathbf{(*;1;*;*;*;*;*;*;*;*)} \leftrightarrow \mathbf{(*;1;*;*;*;2;0;*;*;*)}$$

ce qui suggère qu'au moins 20% des liens du réseau de collaborations de 2004 connectent des auteurs qui ont 1 unique papier à la conférence (valeur 1 sur l'attribut 2), à des auteurs qui ont également 1 papier et qui sont en moyenne 2e dans la liste des auteurs sans avoir été premier auteur sur leurs papiers (valeurs 1, 2 et 0 sur les attributs 2, 5 et 6). La notation '$*$' signifie que l'attribut peut prendre n'importe quelle valeur sur ce schéma.

4 Évolution des sentiments et de la visibilité sur Twitter

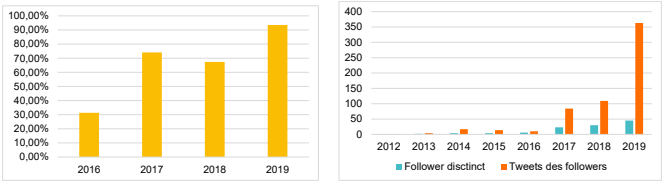
Enfin, nous nous sommes intéressés au rayonnement et à l'appréciation de la conférence et de l'association EGC à travers le média social Twitter. Dans un premier temps, nous avons analysé le jeu de données fourni dans le cadre du défi contenant l'ensemble des messages diffusés (tweets et retweets) par l'association EGC depuis la création de son compte sur Twitter en juin 2016. En plus du texte du message, ce jeu de données contient un ensemble d'information tel que la date de diffusion du message, le nombre de retweets du message, le nombre de j'aime du message, l'utilisateur à l'origine du message (s'il s'agit d'un retweet), le nombre de commentaires, etc. En observant l'évolution du nombre moyen de messages diffusés par mois par l'association EGC, nous avons remarqué que l'activité de diffusion de contenu semble constante au cours des trois dernières années avec une moyenne de messages postés entre 15 et 20 entre 2017 et 2019 (voir figure 4 (a)).



(a) (b)

FIG. 4 – *Évolution du nombre moyen de messages postés (a) et du nombre moyen de retweets et de j'aime par messages (b)*

Puis, nous nous sommes focalisés sur la visibilité des messages postés par l'association EGC, nous avons constaté que ces messages sont en moyenne retweeté au moins une fois et reçoivent au moins un j'aime. De plus, nous avons observé que l'engouement pour les informations diffusées par l'association a augmenté en 2019 avec une moyenne de retweet et de j'aime par tweets supérieurs à 2,5 (voir figure 4 (b)).

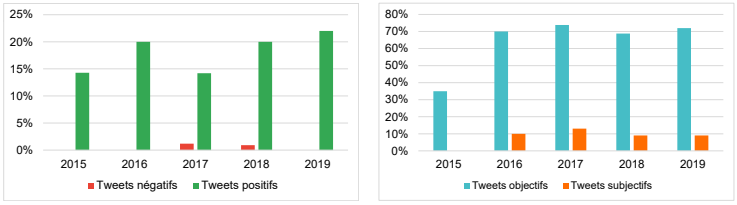


(a) (b)

FIG. 5 – *Évolution de la part de retweets des tweets de l'association parmi ces followers (a) et du nombre de messages postés par les followers relatifs à EGC (b)*

Dans un troisième temps, nous avons enrichi le jeu de données initial en extrayant la liste des followers de l'association EGC ainsi que l'ensemble des tweets qu'ils ont postés depuis la création de leur compte en utilisant l'API de Twitter. Ainsi, nous avons constaté que la part des messages retweetés par les followers avait augmenté entre 2016 et 2019 passants de 31%

à 92% (voir figure 5 (a)). De plus, en analysant les tweets postés par la communauté proche du compte twitter de l'association, nous avons remarqué que de plus en plus de followers diffusent des messages relatifs à l'association ou à la conférence EGC (voir figure 5 (b)).



(a) (b)

FIG. 6 – *Évolution de la part de messages positifs et négatifs (a) et de la part de messages objectifs et subjectifs (b)*

Enfin, nous nous sommes intéressés aux sentiments exprimés par les followers de l'association dans leur tweets relatifs à EGC. Pour ce faire, nous avons utilisé l'API TextBlob [1] qui à partir d'un texte génère un score de polarité et un score de subjectivité. Nous avons observé qu'il y a plus de messages postés marqués positivement que négativement par les followers et qu'il y a une plus grande part de messages objectifs (voir figure 6). La table 1 présente les tweets les plus marqués positivement et négativement par année. Nous observons que la conférence EGC est bien perçue sur Twitter.

TAB. 1 – *Les tweets les plus marqués positivement et négativement par année.*

Année	Tweets	Polarité
2019	Merveilleuse experience @associationEGC 2019 à Metz ! Présentation de nos travaux sur l'augmentation des données pour la classification des séries temporelles.	Positive
2018	Excellente Minute de folie organisée par @guywhiz et Bruno Pinaud à #EGC2018 @associationEGC ! Un sketch de #Coluche dans une conférence c'est pas commun ! ;-) Bravo ! ! !	Positive
2017	Délicieux cocktail avant le dîner de gala à #EGC2017 sous la musique d'un Piano des Gones au @stadedesalpes !	Positive
2016	EGC concurrencé par la European Gas Conference #egc2016 #desambiguisation bon début de conf à tous !	Positive
2015	The slides about the discovery of exceptional local models #olfaction #EGC2015	Positive
2017	#EGC2017 Le poster du pauvre ...	Négative

1. http://textblob.readthedocs.io/en/dev/

5 Conclusion

Dans ce travail, nous avons analysé des données issues de la conférence EGC afin de mettre en exergue depuis 2004, les grandes évolutions de la conférence à travers trois axes : (i) les termes abordés, (ii) les schémas collaboratifs, (iii) les sentiments et la visibilité sur Twitter. Dans un premier temps, nous avons mis en évidence l'évolution des mots les plus fréquents au niveau des titres des articles acceptés chaque année à la conférence. Puis, nous avons généré pour chaque année un réseau de collaboration des auteurs reposant sur plusieurs attributs. Nous avons extrait de ces réseaux les liens fréquents entre les groupes d'auteurs en utilisant une technique de clustering de liens afin d'observer l'évolution des différents schémas collaboratifs. Enfin, après avoir enrichi le jeu de données fourni sur la diffusion de messages de l'association EGC sur Twitter, nous avons observé la visibilité de la conférence sur ce média social ainsi que la polarité et la subjectivité des messages postés par sa communauté proche.

Références

Barrat, A., C. Cattuto, A. E. Tozzi, P. Vanhems, et N. Voirin (2014). Measuring contact patterns with wearable sensors : methods, data characteristics and applications to data-driven simulations of infectious diseases. *Clinical Microbiology and Infection 20*(1), 10–16.

Chung, J. E. et E. Mustafaraj (2011). Can collective sentiment expressed on twitter predict political elections ? In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Sakaki, T., M. Okazaki, et Y. Matsuo (2010). Earthquake shakes twitter users : real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM.

Stattner, E. et M. Collard (2012). Social-based conceptual links : Conceptual analysis applied to social networks. In *International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012, Istanbul, Turkey, 26-29 August 2012*, pp. 25–29.

Van Der Aalst, W. (2016). Data science in action. In *Process Mining*, pp. 3–23. Springer.

Waller, M. A. et S. E. Fawcett (2013). Data science, predictive analytics, and big data : a revolution that will transform supply chain design and management. *Journal of Business Logistics 34*(2), 77–84.

Summary

In this article, we conduct a data analysis work from the EGC conference. These data includes articles accepted at the EGC since 2004 as well as posts posted on the social network Twitter. Our goal is to highlight, since 2004, the major evolutions of the EGC conference in (i) keywords, (ii) collaborations and (iii) sentiments expressed on Twitter. The results obtained highlight the evolution of the keywords of the conference since 2004, the collaborative schemes that are being set up and their evolution, and the sentiments that emerge from the messages published around the conference on Twitter.