Défi EGC 2016 : Analyse par Motifs Fréquents et Topic Modeling

Julien Aligon*, Fabrice Guillet*, Julien Blanchard*, Fabien Picarougne*

*Université de Nantes
Laboratoire Informatique de Nantes Atlantique (LINA)
Equipe Duke
prenom.nom@univ-nantes.fr

Résumé. Dans le domaine de l'analyse de textes, l'extraction de motifs est une technique très populaire pour mettre en évidence des relations fréquentes entre les mots. De même, les techniques de topic modeling ont largement fait leurs preuves lorsqu'il s'agit de classer automatiquement des ensembles de textes partageant des thématiques similaires. Ainsi, ce papier a pour ambition de montrer l'intérêt de l'utilisation conjointe de ces deux techniques afin de mettre en évidence, sous la forme d'un graphe biparti, des mots partageant des thématiques similaires mais aussi leurs relations fréquentes, intra et inter thématiques. Les données du Défi EGC 2016 permettent de valider l'intérêt de l'approche, tout en montrant l'évolution des thématiques et des mots clés parmi les papiers de la conférence EGC sur ces onze dernières années.

1 Introduction

La fouille de données (Han et Kamber (2000)) est devenue un domaine incontournable pour l'analyse de grands volumes de données, auxquels nous sommes désormais confrontés au quotidien (notamment dans le contexte du web). Le principe général de la fouille est d'extraire de l'information pertinente dans l'objectif de caractériser des phénomènes que l'on suppose présents dans les données à traiter. Dans le domaine de l'analyse de textes, l'extraction de motifs fréquents est une technique très populaire pour mettre en évidence des relations fréquentes entre les mots à analyser. De même, les techniques de topic modeling ont largement fait leurs preuves lorsqu'il s'agit de classer automatiquement des ensembles de textes partageant des thématiques similaires. Ainsi, ce papier a pour ambition de montrer la complémentarité et l'intérêt de l'utilisation conjointe de ces deux techniques, afin de produire des visualisations mettant en relation temporellement puis spatialement les thématiques. La visualisation temporelle est basée sur un diagramme de Sankey et permet d'étudier l'influence des thématiques entre elles sur des périodes de temps séquentielles. L'analyse spatiale est construite sur des thématiques et des associations de mots extraits sur un aggloméra de périodes temporelles et prend la forme d'un graphe biparti reliant ces thématiques aux mots. La relation mot-thématique est construite sur une notion de similarité mais aussi sur leurs relations fréquentes, intra et inter thématiques.

Ces deux méthodes sont appliquées aux données fournies par le Défi EGC 2016, et limitées aux résumés des papiers de la conférence EGC sur ces dix dernières années. Le choix de ne

prendre en compte que les résumés s'explique par le fait que l'intégralité des papiers n'est pas disponible pour toutes les années (l'étude temporelle aurait été limitée). De plus, les résumés offrent un vocabulaire, certes restreint, mais très précis. Ce qui en facilite aussi l'exploitation par des techniques de topic-modeling, par exemple.

Le reste de ce papier est organisé comme suit. La section 2 motive l'approche à l'aide d'un exemple simple. Un état de l'art et les principes d'extraction de motifs et de topics sont exposés dans la section 3. La section 4 décrit notre méthode de construction de graphe biparti. Enfin, nous appliquons nos techniques aux données du Défi EGC 2016 et exposons les résultats obtenus dans la section 5.

2 Motivation

Dans cette section, nous illustrons l'intérêt de l'utilisation jointe des méthodes de topic modeling et d'extraction de motifs fréquents, à l'aide de l'exemple ci-dessous.

Considérons les cinq phrases suivantes :

1. Je préfère le café au thé.
2. Je prends trois tasses de café par jour.
3. Il parait que boire du café est bon pour la santé.
4. L'abus de café peut aussi nuire à la santé.
5. Faire du sport améliore le rythme cardiaque et respiratoire.

Si l'on souhaite analyser les tendances générales sur ces phrases, on remarque que, intuitivement, deux thématiques principales émergent :

– les *boissons chaudes*, incluant notamment les mots {*café, thé, tasses*}
– la *santé*, incluant notamment les mots {*santé, rythme cardiaque, respiratoire*}

Pour découvrir automatiquement ce type de thématiques, les techniques de topic modeling sont très appropriées. Par exemple, LDA (Latent Dirichlet Allocation, Blei et al. (2003)) en est un modèle populaire. Mais ce type de modèle ne permet pas d'identifier et caractériser les relations pertinentes entre les mots. En effet, comme on peut le remarquer à l'aide des phrases (3) et (4), il existe une relation fréquente (2/5) entre les mots *café* et *santé*. Ce type de relation aurait été très difficile à détecter sans l'utilisation de techniques de découverte de motifs fréquents. L'utilisation conjointe des techniques de topic modeling et de découverte de motifs peuvent apporter une plus value intéressante pour dégager des tendances dans les textes, à la fois en termes de thématiques mais aussi de relations fréquentes.

3 Extraction de Motifs et Topic Modeling

3.1 Travaux Existants

Il existe peu de travaux combinant à la fois des techniques de topic modeling et de découverte de motifs. On peut notamment citer le travail de Kim et al. (2012), mais dont le but est d'améliorer la pertinence du modèle de découverte des thématiques en y incluant la fouille de motifs fréquents. Cela diffère du but de l'approche de ce papier qui se positionne sur une utilisation conjointe de deux techniques (et non l'une servant à améliorer l'autre). Quelques

travaux (Krestel et al. (2009), Chen et al. (2009)) ont également été réalisés sur la comparaison entre les techniques de topic modeling et motifs fréquents, notamment dans le domaine de la recommandation. On peut enfin citer un papier (Chikhaoui et al. (2012)) qui combine les méthodes de LDA et d'extraction de séquences fréquentes (qui peut être considéré comme un cas plus contraint de l'extraction de motifs).

3.2 Principes théoriques

Dans la suite de ce papier, nous considérons les définitions de base de données, transactions et items :

Definition 1 (Base de données, Transactions et Items)
Soit $D = \{t_1, t_2, ..., t_n\}$ une base de données contenant un ensemble de transactions. Une transaction $t_i = \{i_1, i_2, ..., i_m\}$ contient un ensemble de valeurs d'attributs i_j appelées $item$.

3.2.1 Extraction de Motifs

Le principe d'extraction de motifs fréquents a été introduit par Agrawal et al. (1993). Les motifs permettent d'identifier, à partir d'une base de données, les éléments trouvés fréquemment ensemble. Afin de mesurer la pertinence des motifs fréquents, c'est à dire identifier à quel point le motif est significatif vis à vis des transactions, deux mesures sont utilisées classiquement : les mesures de $support$ et $lift$ chacune définie ci-dessous.

Definition 2 (Support)
La mesure de $support$, noté $supp(X)$ calcule le nombre de transactions parmi D, incluant l'itemset X. Formellement, $supp(X) = |\{t_i \in D, X \subseteq t_i\}|$. La valeur de support est généralement normalisée par la cardinalité de la base de données considérée.

Une autre mesure, utilisée généralement en complément des mesures de support, est le lift. Cette mesure est également définie ci-dessous.

Definition 3 (Lift)
La mesure de $lift$, noté $lift(X, Y)$, permet d'estimer le degré de dépendance entre les itemsets X et Y. L'idée est de pouvoir vérifier que les itemsets X et Y sont bien corrélés entre eux vis à vis de toutes les transactions de la base de données.
Formellement, $lift(X, Y) = \frac{supp(X \cup Y)}{supp(X) \times supp(Y)}$.

Dans le but d'identifier automatiquement les itemsets fréquents, ceux-ci sont extraits à partir d'un ensemble initial de transactions. L'algorithme $Apriori$ (Agrawal et al. (1993)) est une approche classique dans ce contexte, bien que très couteux, en fouillant tous les itemsets fréquents dont la valeur de support est supérieure à un seuil donné.

3.2.2 Latent Dirichlet Allocation

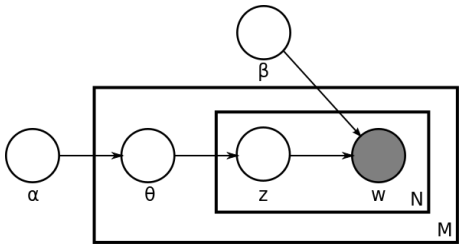
Le topic modeling (Blei et al. (2003)) est un modèle probabiliste permettant de dégager un ensemble de thématiques (ou topics), caractérisant un ensemble de mots, à partir d'un corpus de documents. L'intuition est alors de découvrir les topics les plus probables (vis à vis

des documents) regroupant les mots les plus probables. Une implémentation possible de topics modeling est LDA (Blei (2012)). Afin d'associer des mots à des topics, que l'on supposera découverts selon une probabilité d'apparition, LDA assure que chaque document est caractérisé par un ensemble de topics et que chaque mot fait partie d'un ou plusieurs topics. Plus précisément, LDA est un modèle génératif probabiliste à variables latentes. Les paramètres de ce modèle sont le nombre k de topics à extraire, ainsi que deux paramètres de distributions, α et β. Le paramètre α agit sur la répartition des documents entre les topics et le paramètre β sur la répartition des mots entre les topics.

La représentation graphique du modèle probabiliste LDA est donnée dans la Figure 1 (en *plate notation*).

L'idée de ce modèle est résumé de la manière suivante. On suppose, préalablement, un tirage de k lois de distributions de topics. Pour chaque document $m \in M$, choisir une loi de distribution θ de m parmi les topics. Puis, pour chaque mot $w \in W$ de m, choisir un topic $z \in Z$ respectant la loi θ correspondante.

FIG. 1: Modèle graphique de LDA (repris de Blei et al. (2003))



La principale difficulté pour l'obtention des topics est la complexité exponentielle du calcul de probabilité d'apparition de n'importe quel mots du corpus avec n'importe quel topic. Pour palier ce problème, l'une des heuristiques les plus populaires est basée sur la technique d'échantillonnage de Gibbs.

4 Méthode applicative et représentation visuelle

4.1 Extraction des motifs fréquents

L'extraction des motifs fréquents sur les données EGC va permettre de mettre en évidence les régularités, fréquentes, entre les mots d'une phrase. Les motifs extraits sont composés d'un ou deux items et contraints selon des valeurs minimales de $support$ et de $lift$ (voir Définitions 2 et 3). Le fait de choisir des motifs à un seul item permet de mettre en évidence l'importance, en terme de fréquence, de cet item. Quant aux motifs à deux éléments, ce choix permet d'analyser finement les cooccurrences de mots.

4.2 Extraction des topic model

L'extraction des topic model sur les données du Défi EGC 2016 permet de réaliser automatiquement des regroupements de mots ayant une sémantique proche.

L'implémentation utilisée pour réaliser ces groupes, et basée sur LDA (voir Section 3.2.2), provient de l'API $Mallet$ (McCallum (2002)). Cette implémentation repose sur l'approximation de l'échantillonnage de Gibbs. Ainsi, deux paramètres sont nécessaires pour initialiser l'algorithme :

- α en rapport avec la loi de probabilité des topics vers les documents.
- β en lien avec la loi des mots envers les topics.

Ces lois sont ensuite affinées après n itérations de l'algorithme, jusqu'à obtenir une convergence dans les mots associés aux topics dont le nombre est fixé apriori.

4.3 Représentation visuelle par graphe de voisinage

Afin de faciliter l'analyse des motifs fréquents et de topics extraits selon les principes décrits précédemment, ceux-ci sont organisés sous la forme d'un graphe non dirigé. Ce graphe est construit selon la Définition 4.

Definition 4 (Graphe des Topics et Motifs)
Soit G un graphe biparti non orienté, composé de deux ensembles U et V.
Soit M_i, l'ensemble des mots se rapportant à un topic t_i.
Soit R, l'ensemble des motifs dont l'extraction est conditionnée à des mesures de support et lift. (voir les Définitions 2 et 3).

U représente l'ensemble des noeuds se rapportant aux mots, i.e.
$\forall u \in U, val(u) \in M$ ou $val(u) \in R$, avec val(x), le mot du noeud x.

V représente l'ensemble des noeuds se rapportant aux topics, i.e. $\forall v \in V, val(v) \in T$, $avec\, T = t_1, t_n$ où n est le nombre total de topic et $val(x)$ est le nom du topic.

Les arêtes entre les noeuds de U et V représentent les mots de U se rapportant aux topics de V (voir Section 3.2.2).

Les arêtes entre les noeuds de U représentent les motifs trouvées parmi R.

Pour plus de clarté, un algorithme de positionnement des noeuds est également utilisé. Cet algorithme est basé sur le principe de forces d'attraction/répulsion dont l'implémentation utilisée est $ForceAtlas$ Jacomy et al. (2014).

Example 1 *Supposons que l'on ait à disposition les motifs suivantes :*

- {*café, thé*}
- {*café, santé*}

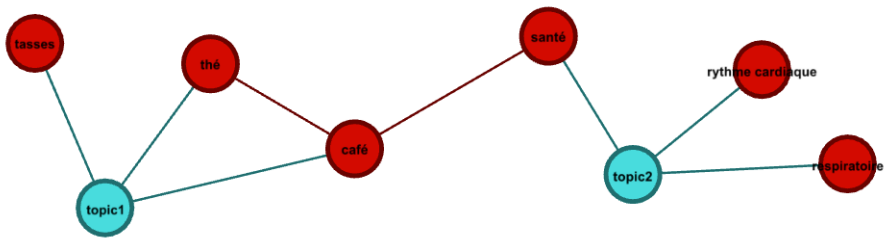
Ainsi que les topics suivants :

- $topic_1$ = {*café, thé, tasses*}
- $topic_2$ = {*santé, rythme cardiaque, respiration*}

La Figure 2 donne un exemple de représentation de ces motifs et topics sous la forme de graphe.

Toute la difficulté de cette représentation est de trouver un bon compromis entre le nombre de topics à extraire (et leur nombre de mots par topics) et le nombre de motifs selon des seuils de support et de lift. En effet, un nombre de topics trop élevé conduirait à rendre le graphe difficile à interpréter. De même, un nombre de motifs trop important associerait trop de noeuds entre eux. A l'inverse, considérer un nombre de topics ou motifs peu important rendrait l'analyse peu informative et donc peu pertinente.

FIG. 2: Graphe de voisinage des topics et motifs.



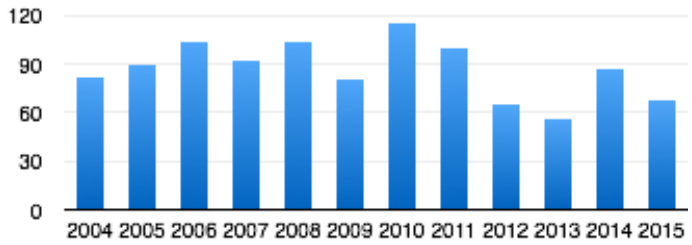
5 Application aux données du Défi EGC 2016

5.1 Préparation des Données

5.1.1 Extraction et Qualité des données

L'extraction des données est basée sur 1041 papiers EGC, entre les années 2004 et 2015. Dans le cadre du défi EGC 2016, nous traitons seulement les résumés des articles, sans prise en compte de la bibliographie. Le nombre de papiers, par année, est représenté Figure 3. On peut, d'ailleurs, y noter un creux particulièrement fort pour les années 2012, 2013 et 2015.

FIG. 3: Fréquences des papiers EGC



Parmi ces papiers, une minorité est rédigée en anglais. Mais, afin de minimiser une possible pollution des données à traiter, ces papiers ont été filtrés lorqu'ils incluaient des mots spécifiques, que l'on retrouve généralement dans tout texte (exemple : the, is, $this$, $paper$). Après filtrage, 80 papiers ont été écartés. A partir du texte d'un résumé particulier, chaque phrase est considérée comme formant une transaction (voir Définition 1). Ces transactions sont ensuite nettoyées de l'éventuelle présence de caractères non gérés par les algorithmes de fouille de données. Un total de 4007 transactions est ainsi produit à partir des données du défi.

5.1.2 Pos-tagging et Lemmatisation des transactions

En considérant l'ensemble des transactions obtenues lors de la phase d'extraction des données, une phase d'étiquetage morpho-syntaxique et de lemmatisation est appliquée. Cette phase permet, en effet, de grandement simplifier le travail de fouille et l'analyse des résul-

tats, en considérant les formes canoniques des mots (verbes à l'infinitif, noms au singulier par exemple). L'étiquetage est réalisé à l'aide de l'outil TreeTagger (Tre), permettant d'annoter automatiquement chaque mot présent dans chacune des transactions. Dans un premier temps, seuls les mots de type verbes, noms communs et noms propres ont été considérés. Plus précisément, les lemmes de chacun de ces types sont pris en compte, afin de faciliter l'analyse des résultats fournis par les algorithmes de fouille. L'analyse plus complète des autres types de mots (adjectifs, nombres, etc.) fait partie des perspectives d'amélioration. Un certain nombre de lemmes ont également été filtrés, car considérés comme non pertinents. En effet, les lemmes présentant une grande fréquence d'apparition dans les résumés (par exemple les verbes *être* et *avoir*) ne peuvent que contribuer à bruiter l'analyse. De même, des mots trop génériques du type *papier*, *sembler*, *présenter*, *prendre*, etc. ajoutent aussi clairement une pollution inutile. A titre d'indication, nous donnons la fréquence d'apparition de chacun des types de mots considérés (les doublons de mots sont comptabilisés), après filtrage, sur l'ensemble des résumés. On remarque ainsi qu'il y a presque 3 fois moins de verbes à traiter (10 619) que de noms communs (29 471) et que très peu de noms propres sont présents (10).

5.2 Résultats obtenus

Cette section discute des tests et des résultats obtenus après analyse des papiers EGC sur ces onze dernières années. Dans un premier temps, l'étude se focalise sur la seule extraction de topics, où sont décrits les paramètres utilisés dans l'algorithme LDA, ainsi qu'une description des thématiques extraites pour chaque année de la conférence, notamment à l'aide d'un diagramme de Sankey. Dans un second temps, une analyse du graphe biparti, contenant à la fois les topics et les motifs extraits sur l'ensemble du corpus, est développée.

5.2.1 Etude des topics

Cette section détaille les résultats obtenus quant à l'évolution des topics extraits pour chaque année de la conférence EGC, de 2004 à 2015. L'implémentation LDA utilisée provient de l'API $Mallet$ (McCallum (2002)). Le paramètre β (voir Section 4.2) est calculé à l'aide de l'optimiseur proposé par Mallet. Le paramètre α est fixée à 0.1, en supposant que les résumés EGC se focalisent sur un petit nombre de topics. Le nombre d'itération est fixé à 2000. Le nombre de topics à extraire est fixé à 10, par année, et le nombre de mots par topic à 10. Ces valeurs, pouvant être considérées comme plutôt faibles, permettent de faciliter l'interprétation manuelle des résultats pour l'analyse des papiers EGC. Afin de suivre l'évolution des topics, une mesure de similarité entre topics est également définie. Cette mesure est basée sur la similarité cosinus. Formellement, la similarité entre deux topics est définie comme suit : $sim(t_1, t_2) = cos(V_1, V_2)$ où V_1 et V_2 sont des vecteur des probabilités normalisées d'apparition des mots dans les topics t_1 et t_2, respectivement. L'évolution des topics similaires est représentée à l'aide d'un diagramme de Sankey (à l'image de ce qui a été réalisé dans Pépin et al. (2015)). Ainsi, les flux de ce diagramme représentent les proximités relatives d'un topic par rapport à un autre, entre deux années successives. Le diagramme sur les papiers EGC peut être retrouvé Figures 4a et 4b. Pour une meilleure visibilité, l'ensemble du diagramme peut être consulté sur le site suivant : `https://www.dropbox.com/sh/ozjwezc2wu6s5sq/AADD-515AUHnFnuCMvOX-vtfa?dl=0` L'épaisseur des flux entre topics représente la

valeur de similarité cosinus. La longueur des topics dépend du nombre de flux sortant et entrant, ainsi que de leurs valeurs de similarités. Cela permet, visuellement parlant, de constater rapidement quels sont les topics agrégeant le plus de liens provenant de différents topics des années précédentes. Les mots associées aux topics représentent trois mots (pour des raisons de lisibilité) ayant le plus de probabilité d'appartenir au topic considéré. Le seuil de similarité cosinus est fixé à 20%. Ce seuil permet d'assurer une visibilité du plus grand nombre de topics, tout en évitant un croisement de flux trop nombreux, ce qui rendrait l'analyse difficile. Le diagramme offre aussi la possibilité de représenter les topics qui ne sont similaires avec aucun autre (appelés *singleton*) (non représentés dans les figures, par manque de place). Le nombre de ces singletons évolue peu entre les années (entre 2 et 3 par année). A noter que l'augmentation du nombre de ces singleton pour l'année 2015 est attendu, puisque s'agissant de la dernière année référencée.

De manière globale, nous constatons sur ce schéma et dans le tableau 1 que le nombre de flux sortant par topic diminue de 2004 à 2011 pour se stabiliser à un seul flux sortant par topic. De manière symétrique, le nombre de flux entrant suit la même tendance. Cette représentation nous apprend que les topics se sont petit à petit structurés et figés au fur et à mesure des années dans la conférence EGC. Nous pouvons interpréter cette tendance comme une maturation des thématiques qui sont maintenant bien établies. Par exemple, entre les années 2004 et 2005, des thématiques assez générales sont identifiées : analyse de documents, techniques d'apprentissage et de classification, fouille de règles d'association. Au contraire, pour les années 2014 et 2015, des thématiques plus spécifiques comme la fouille de motifs dans les images, les partitions, les mesures de proximité ou la recommandation font leur apparition. Cela ne signifie pas que ces thématiques n'étaient pas déjà présentes au fur et à mesure des années, mais elles apparaissent bien plus clairement sur les années récentes.

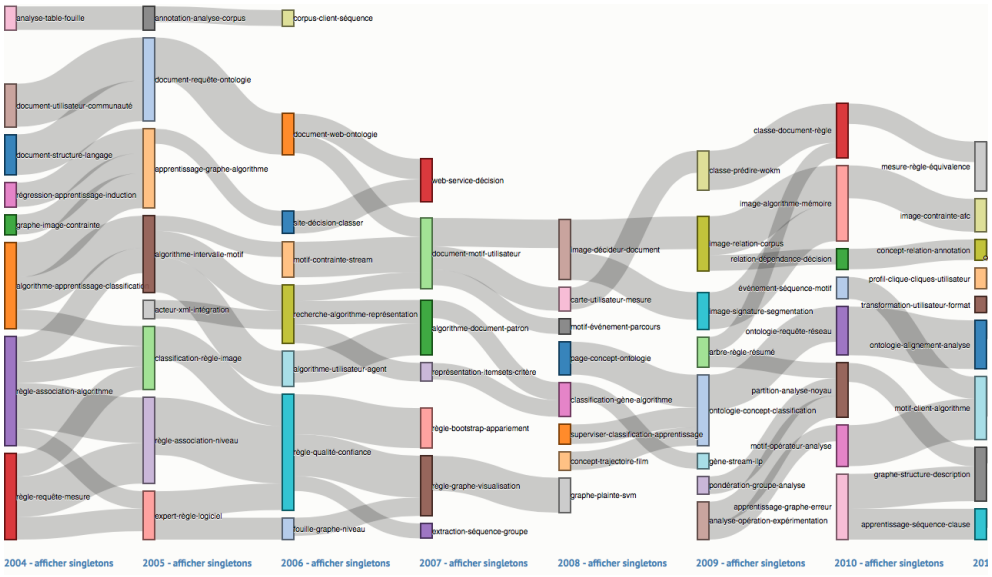
flux	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
entrant	1,71	1,40	1,17	1,50	1,33	1,00	1,14	1,00	1,00	1,00	1,00	
sortant		1,71	1,40	1,40	1,00	1,00	1,25	1,14	1,00	1,00	1,33	1

TAB. 1: Nombre moyen de flux entrants et sortants d'un topic par année. On ne prend ici en compte que les topics ayant au moins un flux sortant.

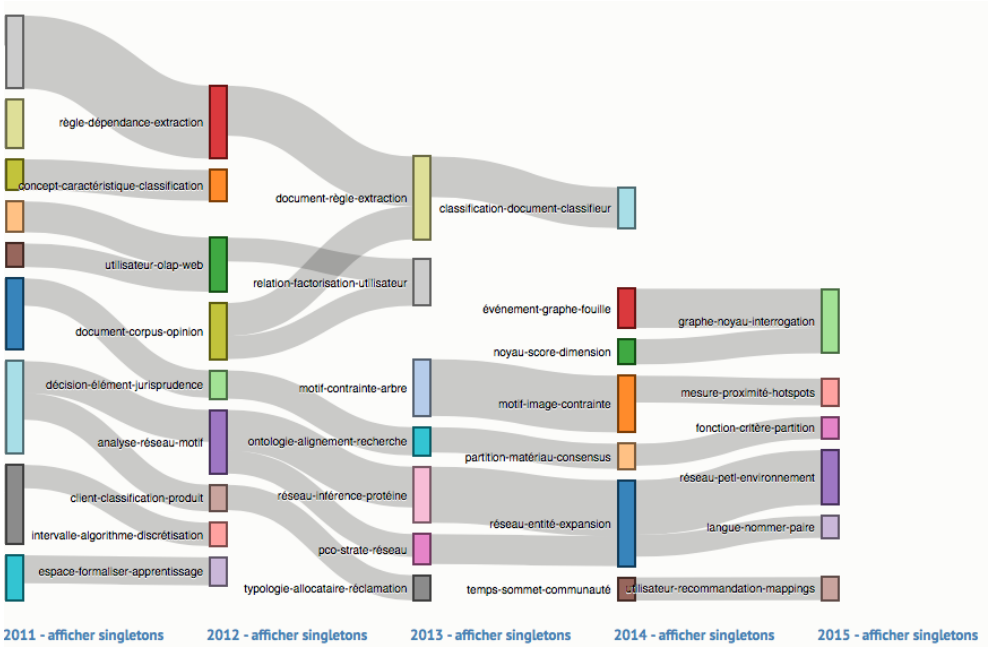
5.2.2 Etude du graphe des motifs et topics

Cette section rend compte des résultats obtenus par le graphe biparti (voir Section 4.3) incluant à la fois une représentation des motifs fréquents et des mots des topics. L'analyse précédente s'est focalisé sur la temporalité des topics. Nous avons constaté une certaine stabilité des thématiques EGC au cours des dernières années. Ainsi, les topics et les motifs du graphe biparti sont extraits sur l'ensemble du corpus. Cela permet de comparer les topics stables entre eux à l'aide des relations fréquentes entre les mots.

L'extraction de motifs repose sur deux seuils (voir Section 3.2.1) : 1) le support minimal, fixé à 12 transactions et 2) le seuil de lift, fixé à 2. Ce dernier seuil peut être considéré comme assez élevé mais il permet de s'assurer que les mots des motifs soient clairement dépendant l'un de l'autre. Les résultats sont présentés Figure 5. Pour une meilleure lisibilité, l'ensemble du graphe biparti est disponible à cette adresse : `https://www.dropbox.com/`



(a) Années 2004 à 2011



(b) Années 2011 à 2015

FIG. 4: Diagramme de Sankey pour les données du Défi EGC de 2004 à 2015.

sh/ozjwezc2wu6s5sq/AADD-515AUHnFnuCMvOX-vtfa?dl=0 Pour plus de visibilité, les liaisons entre mots extraits par les motifs sont colorés en bleu clair. Les liaisons entre mots et topics sont en gris clair. L'épaisseur des mots correspond dépend du nombre de liens entrants et sortants. L'épaisseur des liaisons correspond à la valeur de support vis à vis du topic ou du mot. Les caractéristiques de chacun des topics de la Figure 5 peuvent être résumés comme ceci : le topic 1 traite de la détection de trajectoires ; le topic 2 des graphes et réseaux ; le topic 3 des ontologies et documents ; le topic 4 de la classification et apprentissage ; le topic 5 des motifs et séquences ; le topic 6 des mesures de similarité ; le topic 7 des entrepôts et OLAP ; le topic 8 de l'analyse de textes ; le topic 9 des règles d'association ; le topic 10 de l'analyse d'image.

On peut noter sur le graphe produit que les topics centraux semblent être les topics 5 et 2 (et dans une moindre mesure, le topic 3). En effet, l'extraction de motifs montre clairement de nombreux liens entre les mots de ces topics vis à vis des autres topics. Certains mots sont également bien identifiés comme centraux dans le graphe. Comme attendu, les mots *extraction*, *analyse*, *document* et *classification* sont au coeur des motifs extraits. A noter tout de même, et de manière surprenante, que le mot *qualité* n'est pas mise en évidence dans les motifs. Les termes *motifs* et *règles d'association* sont aussi clairement différenciés. Intuitivement, on peut supposer que le terme *motif* est plus générique que le terme *règle d'association* et s'utilise ainsi plus facilement pour des papiers traitant de la fouille de données en général.

6 Conclusion et Perspectives

Nous proposons dans cet article de montrer la complémentarité et l'intérêt de l'utilisation conjointe des techniques d'extraction de motifs fréquents et de topic modeling. Les visualisations temporelles entre thématiques ont permis de mettre en évidence une stabilité des thématiques, notamment sur les dernières années de la conférence EGC. L'analyse spatiale des thématiques, réalisée à l'aide d'une graphe biparti associant découverte de topics et de motifs fréquents, ont permis de rendre compte des relations existantes entre thématiques. En perspective de travail, il serait intéressant de produire la même analyse de papiers sur d'autres conférences pour déterminer s'il est possible d'établir un cycle de vie particulier attaché aux conférences, caractéristique du degré de maturation de celles-ci, voire même de caractériser et de prédire leurs évolutions. Le graphe biparti pourrait aussi tout à fait être utilisé dans le cadre de la recommandation (appliquée au texte), permettant d'exploiter à la fois des relations fréquentes entre mots et leurs thématiques.

Références

Logiciel *treetagger*. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger.

Agrawal, R., T. Imieliński, et A. Swami (1993). Mining association rules between sets of items in large databases. *SIGMOD Rec. 22*(2), 207–216.

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.

Chen, W.-Y., J.-C. Chu, J. Luan, H. Bai, Y. Wang, et E. Y. Chang (2009). Collaborative filtering for orkut communities : Discovery of user latent behavior. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, New York, NY, USA, pp. 681–690. ACM.

Chikhaoui, B., S. Wang, et H. Pigot (2012). Adr-splda : Activity discovery and recognition by combining sequential patterns and latent dirichlet allocation. *Pervasive Mob. Comput. 8*(6), 845–862.

Han, J. et M. Kamber (2000). *Data Mining : Concepts and Techniques*. Morgan Kaufmann.

Jacomy, M., T. Venturini, S. Heymann, et M. Bastian (2014). Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE 9*(6), e98679.

Kim, H. D., D. H. Park, Y. Lu, et C. Zhai (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology 49*(1), 1–10.

Krestel, R., P. Fankhauser, et W. Nejdl (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the Third ACM Conference on Recommender Systems*, RecSys '09, New York, NY, USA, pp. 61–68. ACM.

McCallum, A. K. (2002). Mallet : A machine learning for language toolkit. `http://www.cs.umass.edu/~mccallum/mallet`.

Pépin, L., N. Greffard, P. Kuntz, J. Blanchard, F. Guillet, et P. Suignard (2015). Visual analytics for exploring the topic evolution of company targeted tweets. In *Proc. of the 45th Int. Conf. on Computers & Industrial Engineering*.

Summary

In the field of text analysis, pattern mining remains a very popular technique for highlighting the frequent relationships between words. Similarly, topic modeling techniques have proven their worth for automatically classifying sets of texts sharing similar topics. Thus, this paper aims to show the benefit of the combined use of these techniques to highlight, as a bipartite graph, words sharing similar topics but also their frequent relations, intra and inter topics. The data of the Défi EGC 2016 are used to validate the interest of the approach, while showing the evolution of topics and keywords among the papers of the EGC conference on the last eleven years.

FIG. 5: Graphe biparti des papiers EGC.