Topic modeling and hypergraph mining to analyze the EGC conference history

Adrien Guille*, Edmundo-Pavel Soriano-Morales*, Ciprian-Octavian Truica**

*Laboratoire ERIC, Université Lumière Lyon 2
{adrien.guille;edmundo.soriano-morales}@univ-lyon2.fr
**Computer Science dept., University Politehnica of Bucharest
ciprian.truica@cs.pub.ro

Abstract. Each year the EGC conference gathers researchers and practitioners from the knowledge discovery and management domain to present their latest advances. This year's edition features an open challenge that encourages participants to leverage the EGC rich anthology which spans from 2004 to 2015. The ultimate goal is to highlight the dynamics of the conference history and to try to get a glimpse of the coming years. In this context, we first describe our methodology for inferring latent topics that pervade this corpus using non-negative matrix factorization. Based on the discovered topics and other properties of the articles (e.g., authors, affiliations) we shed light on interesting facts on both the topical and collaborative structures of the EGC society. Secondly, we employ a hypergraph itemset extraction process to discover existent but latent relations between authors or between topics. We also propose topic-author and author-author recommendations with a content-based approach. Lastly, we describe a Web interface for browsing this collection of articles complemented with the discovered knowledge.

1 Introduction

In this article, we describe work done in the context of the first edition of the EGC challenge, which ultimate goal is to highlight the dynamics of the conference history and to try to get a glimpse of the coming years.

Dataset Participants of the challenge are provided with the descriptions of 1935 articles published by RNTI, out of which 1041 are articles presented at the EGC conference. Each article is described by several fields: year, title, abstract (potentially missing), list of authors, and a URL pointing at the first page of the article (potentially missing or unreachable). When it is possible, we enrich the descriptions of the articles with (i) the language detected from their abstracts and (ii) the authors' affiliations. Language detection relies on a naive Bayes classifier, trained on Wikipedia articles covering French and English, using n-grams of characters as features (Cavnar and Trenkle, 1994). The identification of authors' affiliations relies on regular expressions to match e-mail addresses in the content of the first page of the articles, assuming

TAB. 1: Dataset statistics.

Type of articles	Count (approx. proportion)
Articles for which the abstract is available	896 (86%)
Articles for which the first page is available	936 (89.9%)
Articles for which the detected language is French	817 (78.5%)
Articles for which affiliation are known	893 (86%)

that the domains of these addresses allow identifying institutions. Tab. 1 gives some statistics about this dataset.

Main results and organization of the paper In Sec. 2, we present results based on latent topics discovered from the content of the articles' abstracts. We find that the scope of the EGC conference is evolving through time, notably to incorporate novel interesting issues. We also find that, even though most of the works presented at the conference emanates from the academia, the industry is very active on some targeted issues. Furthermore, we observe that authors have different ways of collaborating inside a topic or across topics, and also that some authors are highly specialized on a topic whereas others are kind of generalists. In Sec. 3, we show the results for two recommendation approaches: hypergraph-based mining and content-based recommendation. We find that the recommendations suggested make sense according to the topics each author works with. We also observe that there are certain research domains that seem to be interesting to develop. Additionally, we show interesting recommendations for highly-published and new-coming authors at EGC. In Sec. 4, we describe a Web interface for exploring the anthology complemented with the extracted knowledge.

2 Topical structure of the EGC conference

The purpose of this section is to shed light on some interesting facts about the EGC conference, based on its topical structure. We begin by describing our methodology for inferring latent topics from the abstracts of the articles, before analyzing these topics in details.

2.1 Methodology

Preparing the articles To avoid language-specific topics, and because EGC is mainly a French-speaking conference, we choose to consider only articles which abstract is written in French. In order to improve the quality of the discovered topics, we lemmatize the abstracts. For that, we use MElt, a maximum entropy Markov model-based part-of-speech tagging system especially designed for French (Denis and Sagot, 2012), and Lefff, a morphological and syntactic lexicon for French (Sagot, 2010), to match {word, part-of-speech} pairs with lemmas. In addition, we prune lemmas which absolute frequency in the corpus is less than 4, as well as lemmas which relative frequency is higher than 80%, with the aim to only keep the most significant ones. Eventually, we build the vector space representation of these articles with $tf \cdot idf$ weighting. It is a $n \times m$ matrix denoted by A, where each line represents an article, with $n = 817$ (*i.e.* the number of articles) and $m = 1739$ (*i.e.* the number of lemmas).

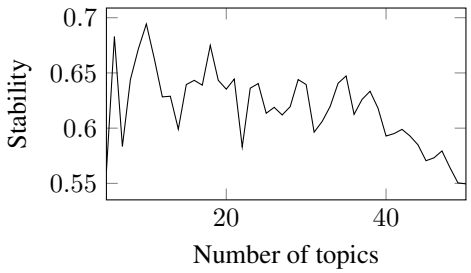

FIG. 1: Weighted Jaccard average stability measure for a number of topics varying between 5 and 50 (higher is better).

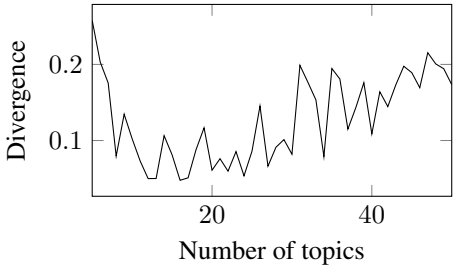

FIG. 2: Symmetric Kullback-Liebler divergence for a number of topics varying between 5 and 50 (lower is better).

Choosing a topic model Given the matrix A and a small number of topics k ($k \ll m$), a topic model consists in two matrices: W and H. W is a $n \times k$ matrix that describes the articles in terms of topics, and H is a $k \times m$ matrix that describes topics in terms of words. More precisely, the coefficient $w_{i,j}$ defines the importance of topic j in article i, and the coefficient $h_{i,j}$ defines the importance of word j in topic i. We consider two candidate methods for topic modeling: (i) Latent Dirichlet Allocation (LDA), a probabilistic generative topic model proposed by Blei et al. (2003), and (ii) Non-negative Matrix Factorization (NMF), a vector space factorization method which has recently became popular for topic modeling (Berry and Browne, 2005). It should be noted that we use variational inference for LDA and a projected gradient method for NMF.

Estimating the optimal number of topics Determining an appropriate value of k is critical to ensure a pertinent analysis of the EGC anthology. If k is too small, then the discovered topics will be too vague; if k is too large, then the discovered topics will be too narrow and may be redundant. To help us with this task, we compute two metrics. First, we compute the metric proposed by Greene et al. (2014), which is based on the assumption that a model with an appropriate number of topics is more robust to missing data. Given a value of k, it assesses the stability of the discovered topics for several models fitted to sub-samples of the original data. The stability is measured in terms of the average weighted Jaccard distance between the ordered sets of words describing the topics. Thus, the higher the stability, the closer k is to a suitable value. We also compute the metric proposed by Arun et al. (2010), which is based on the assumption that the distribution of the singular values of H is close to the distribution of row-L_2 norm of W when k is close to an optimal number of topics. This metric is defined as the symmetric Kullback-Liebler divergence between these two distributions, which means that the lower the divergence is, the closer k is to a suitable number of topics.

2.2 Results

We find that LDA and NMF identify very similar topics in this corpus, with a slight advantage for NMF in terms of topic separability. Guided by the two metrics described previously, we manually evaluate the quality of the topics identified with k varying between 15 and 20.

TAB. 2: Descriptions of the discovered topics.

Topic id	Most relevant words
topic 0	réseau, social, communauté, détection, méthode, analyse, interaction, lien
topic 1	ontologie, alignement, sémantique, annotation, concept, domaine, owl, entre
topic 2	règle, association, extraction, mesure, base, extraire, confiance, indice
topic 3	séquence, temporel, événement, série, modèle, évènement, vidéo, spatio
topic 4	motif, séquentiel, extraction, contrainte, fréquent, extraire, découverte, donnée
topic 5	document, xml, annotation, recherche, information, structure, requête, mots
topic 6	utilisateur, web, information, site, système, page, sémantique, comportement
topic 7	connaissance, gestion, expert, agent, système, compétence, modélisation
topic 8	variable, classification, superviser, méthode, classe, apprentissage, sélection
topic 9	image, afc, segmentation, recherche, région, objet, classification, satellite
topic 10	graphe, voisinage, représentation, interrogation, fouille, visualisation, structure
topic 11	donnée, flux, base, requête, cube, fouille, visualisation, entrepôt
topic 12	algorithme, arbre, svm, ensemble, décision, nouveau, grand, résultat
topic 13	carte, topologique, auto, organisatrice, som, cognitif, probabiliste, contrainte
topic 14	texte, corpus, automatique, textuel, partir, méthode, opinion, clr

For illustration, Fig. 1 and Fig. 2 respectively plot the stability-based and divergence-based metrics for $k \in [5; 50]$ using NMF. Eventually, we judge that the best results are achieved with NMF for $k = 15$. Table 2 lists the most relevant words for each of the 15 topics discovered from the articles with NMF. They reveal that the people who form the EGC society are interested in a wide variety of both theoretical and applied issues. For instance, topics 8 and 12 are related to theoretical issues: topic 8 covers papers about model and variable selection, and topic 12 covers papers that propose new or improved learning algorithms. On the other hand, topics 0 and 6 are related to applied issues: topic 0 covers papers about social network analysis, and topic 6 covers papers about Web usage mining.

In the following, we leverage the discovered topics to highlight interesting particularities about the EGC society. To be able to analyze the topics, supplemented with information about the related papers, we partition the papers into 15 non-overlapping clusters, *i.e.* a cluster per topic. Each article $i \in [0; 1 - n]$ is assigned to the cluster j that corresponds to the topic with the highest weight w_{ij}, as formalized in Eq. 1.

$$cluster_i = \underset{j}{\operatorname{argmax}}(w_{i,j}) \tag{1}$$

Shifting attention, evolving interests Fig. 3 shows the frequency of topics 0 (social network analysis and mining) and 2 (association rule mining) per year, from 2004 until 2015. The frequency of a topic for a given year is defined as the proportion of articles, among those published this year, that belong to the corresponding cluster. This figure reveals two opposite trends: topic 0 is emerging and topic 2 is fading over time. While there was apparently no article about social network analysis in 2004, in 2013, 12% of the articles presented at the conference were related to this topic. In contrast, papers related to association rule mining were the most frequent in 2006 (12%), but their frequency dropped down to as low as 0.2% in

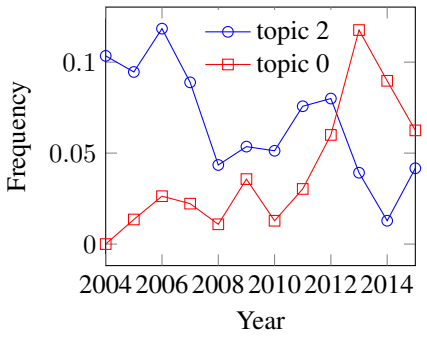

FIG. 3: Frequency of topic 0 (asocial network analysis) and topic 2 (association rule mining) per year.

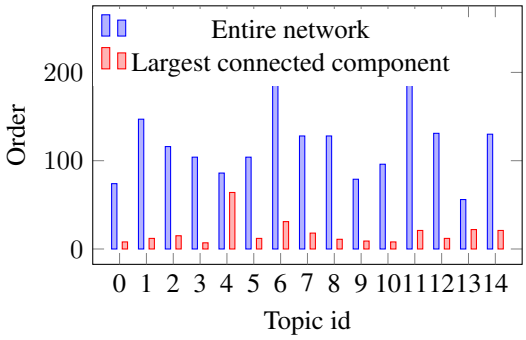

FIG. 4: Order of the collaboration network and order of the largest connected component for each topic.

2014. This illustrates how the attention of the members of the EGC society is shifting between topics through time. This goes on to show that the EGC society is evolving and is enlarging its scope to incorporate works about novel issues.

Collaborations across topics, with few collaborators per topic Fig. 4 shows the order (*i.e.* number of nodes) of the 15 topic-wise collaboration networks and the order of the largest connected component in each of these networks. Given a topic, the nodes of the collaboration network represent the authors of papers assigned to the related cluster. Edges connect pairs of authors who co-wrote one or more papers assigned to this cluster. We observe that all collaboration network are made of lots of small connected component (mostly cliques), except for the collaboration network constructed from papers assigned to topic 4 (pattern mining), which mainly consists in a large connected component. The average proportion of authors in the largest connected component in the topic-wise collaboration networks is about 0.16 (0.74 for topic 4), whereas the proportion of authors in the largest connected component in the global network (*i.e.* made of all the collaborations with no consideration for topics) is about 0.47. This indicates that authors tend to collaborate with different set of authors across different topics.

Mostly academic research, intense industrial research on targeted issues Fig. 5 shows the number of articles, per institution, related to topic 8 (model and variable selection). Whereas every other topic is dominated by academic institutions – in terms of number of publications, this topic is dominated by the industry. More specifically, there are 12 distinct papers involving authors affiliated to Orange (*i.e.* using an e-mail address which domain is either orange-ftgroup.com, orange-ft.com or orange.com).

Specialized authors, generalist authors Fig. 7 and Fig. 8 respectively show the weight distribution over topics for the most specialized author (who has published 9 papers), and the most generalist author (10 papers) among the 88 authors who have at least 5 publications. We quantify the degree of specialization of authors in terms of the Pearson's moment coefficient of skewness of their respective weight distribution over topics. The higher the skewness, the more

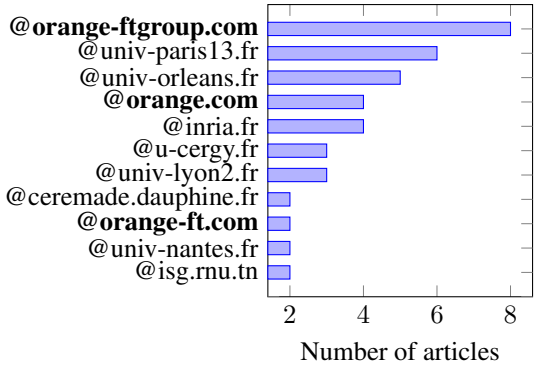

FIG. 5: Number of articles assigned to topic 8 vs. authors' affiliations. Only affiliations with at least 2 publications are shown.

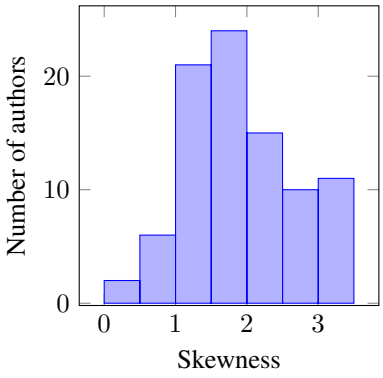

FIG. 6: Histogram of per-author skweness of the weight distribution over topics.

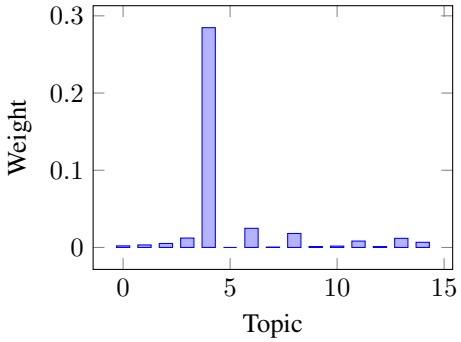

FIG. 7: Weight distribution over topics for the most specialized author.

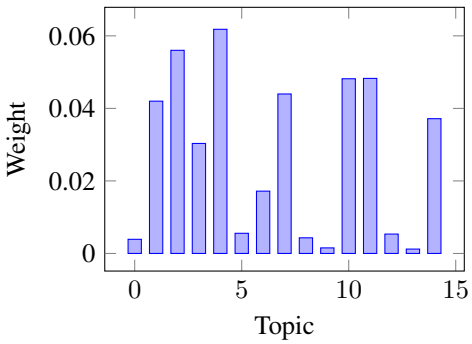

FIG. 8: Weight distribution over topics for the most generalist author.

specialized on a particular topic an author is. For a given author, this distribution is obtained by averaging the weight distributions of all the papers he's published. Fig. 6 shows the histogram of the skweness coefficient measured for the 88 authors mentioned above. It shows that there are few generalist authors, with most of the authors having a skewness between 1 and 2, which roughly corresponds to a distribution with two major topics.

3 Recommendations

The following section describes the methodology we use to propose author-author, topic-topic and author-topic recommendations. We employ two different approaches: on the one hand we consider the interactions between authors as transactions and extract latent frequent itemsets using a hypergraph-based mining method. The itemsets discovered can indeed be seen as recommendations. On the other hand, we directly make recommendations based on

a bipartite graph and using classical item similarity measures. Both approaches are described below.

3.1 Background

Hypergraph-based mining We aim to find interesting relations between authors. To that end, we follow the hypergraph-based item discovery method proposed by Liu et al. (2011). The goal is to find frequent itemsets (between authors or topics) not based on their support (frequency of apparition) but on their similarity as measured by the Average Commute-Time (ACT) distance Fouss et al. (2007). It is our intuition that close authors (related by their topics and years of publication) are good candidates to work together in the future. The ACT distance, a random-walk based vertex distance, has the advantageous property of decreasing when the number of paths connecting two nodes increases and when the length of the paths decreases. In this experiment we work with the complete EGC French corpus in order to possibly find a recommendation for each author in the dataset.

Content-based recommendation This type of recommendation uses attributes to describe items which are then proposed to the users. In our case, the users are the authors and the items are the previously found topics, which are indeed discovered by leveraging the words used in the author's articles. We use an item-similarity recommendation system Ricci et al. (2011) to suggest authors and topics for each author. Briefly, given an author, our model suggests ranked topics according to their similarity to the topics observed for the author in question. Through the similarity between said topics, we can also find the most similar authors for each author.

3.2 Methodology

3.2.1 Hypergraph mining

Filtering authors The first step we follow is to build a list of authors from the EGC French anthology. From the 817 articles, we drew a list of 1307 individual authors. We could have kept only the first appearing authors of each paper, assuming that they contributed most of the writing. Nonetheless, the author name placement is a policy particular to each research laboratory that may or may not represent the amount of contribution to the submitted manuscript. Still, we do filter the authors according to their number of publications.

Creating the hypergraph incidence matrices We want to influence the similarity between a pair of authors according to their shared topics and to their co-occurrent years of publication. Our intuition is that two authors may work together if they share similar topics and if both are active authors during a similar period of time. In order to account for these two characteristics we build two hypergraph incidence matrices. An **author-topic incidence matrix** that links authors with topics, and an **author-year incidence matrix** which describes the relation between authors and their respective articles' years of publication.

The author-topic incidence matrix is created by defining authors as vertices (rows) and topics as hyperedges (columns). Each cell contains 1 if the corresponding author is represented by any of the 15 topics, else it contains 0. In this setting, authors are first represented by their articles and these are depicted by the list of topics found in Section 2. Thus, each author is

indirectly represented by a set of topics. It is clear that an author might not be represented by all the topics found, thus it is important to determine the degree of influence of each topic according to the weight assigned by the topic model. Thus, for a given author and for each of her papers, we first give more importance to those topics with higher weight (or probability) and then we select those topics that appear several times across the author set of papers. The chosen topics are used to represent each author in the author-topic incidence matrix

As said before, we want to influence the similarity between authors by considering the years they have published in. The author-year matrix has authors as vertices (rows) and the EGC French anthology years (from 2004 to 2015) as hyperedges. Each cell contains 1 if the corresponding author published in any of the mentioned years. It contains zero otherwise.

Building the $\mathbf{S}_{CT}$ similarity matrix In order to take into account the information from both incidence matrices described before, we compute a square ACT similarity [1] matrix (with authors as rows and as columns) for each incidence matrix. Finally, both matrices are joined into a single similarity matrix $\mathbf{S}_{CT}$ by averaging both matrices into one.

Finding the frequent 2-itemsets Once the similarity matrix $\mathbf{S}_{CT}$ is computed, we can determine couples of related authors by setting a threshold on the similarity value between each of them. If the threshold is exceeded, that is, two authors have a similarity higher than the fixed threshold, we consider both of them as a frequent authors' itemset and thus a possible co-working recommendation.

Finding the best parameters Measuring the performance of recommendation systems is complex and can be subjective. As a measure of performance for the author-author relation mining, we set to maximize the ratio α defined as the number of real articles written by the mined interesting pairs of authors divided by the total number of pairs found. Our system has three tunable parameters: (i) θ, the minimum similarity value in the S_{CT} matrix to consider a couple of authors as related, (ii) γ, the minimum number of publications an author should have, and (iii) λ, the number of topics that describe each author. To find the best possible α ratio, we start a grid search using the following range of values for each parameter: $\theta = \{0.5, 0.55, 0.6, ..., 0.95\}$, $\gamma = \{3, 4, 5, ..., 9\}$ and $\lambda = \{2, 3, 4, ..., 15\}$.

3.2.2 Content based recommendations

In this approach we want to study two specific groups of authors: the *veteran* and the *newcomer* group. Veteran authors are the top ten researchers with the highest number of publications. Newcomer authors are those researchers that have two or more publications and began participating in the EGC conference since 2012. Our goal is to offer to each researcher in these groups a set of recommendations with whom it would be interesting to collaborate with. For the veteran authors, it may allow them to diversify and enlarge their research domain. For the newcomer writers, a clear map of who is working in similar topics, at a relatively close level, may help to improve their academic careers. At the very least, these recommendations can help make $networking$ easier during the conference itself. We show the recommendations for these authors in the form of recommendation graphs below.

1. Given that ACT is a distance, we transform it into a similarity metric by normalizing and subtracting from one.

Filtering authors and representation We find the authors belonging to the veteran and newcomer groups by applying the filters described above. Each authors is represented by the top five topics (sorted by their weight, as determined by the NMF or LDA approach) that best describe them. At this point, we create a bipartite graph between authors and topics connected by weighted edges (the weights being the same described before).

Recommending authors and topics To recommend topics to authors, our model first builds a similarity matrix between topics using the observations of authors described by their related topics. Then, it scores a topic t for author a using a weighted average over the past observations of a, that is, the topics already used by a. In a similar fashion, the model recommends authors to authors by ranking them according to their similarity using the topics shared between them.

Choosing optimal parameters and best similarity measure The parameters of this technique are two: γ, which, as above, is the minimum number of publications; and λ, the number of topics that describe each author. A similarity measure is needed to relate authors with other authors or with topics. We test three types of similarity measures that : Jaccard, Pearson and cosine similarities. Experimentally, we found that the cosine similarity is the best performing while maximizing the real collaboration's ratio described in the previous approach.

3.3 Results

Hypergraph author-author mined relations After running the grid search to find the optimal parameters for the author-author recommendation, we found the maximal value $\alpha = 0.33$, that is, one third of the found predictions exists already in the EGC anthology. The values of the parameters are: $\theta = 0.65$, $\gamma = 8$ and $\lambda = 2$. In this setting, 37 authors are considered and 9 predictions exceed the θ threshold. The mined relations between authors are shown in Table 3. The three relations already existing in the EGC anthology are shown in bold. According to our experiments, we find that those names are indeed close according to the words shared by their representative topics.

Hypergraph topic-topic mined relations In order to propose topic-topic recommendations, that is, which two topic domains could work together according to the authors involved in both of them, we simply transpose the author-topic matrix (described above in paragraph 3.2.1) and directly calculate the similarity matrix S_{CT}. We use the same parameters found in the previous section, except for the number of topics describing each author, λ, which was set to 5, to avoid a singular matrix during the computations. The results are shown in Table 4. In general, the recommendations suggest the intersection of semantic Web technologies with non-structured data, such as images and text documents.

Recommendations for newcomer and veteran authors In Fig. 9 we can see the top three suggestions for the 10 veteran authors. We keep the recommendations only between the veteran authors themselves. Fig. 10 shows the top three recommendation graph for the newcomer authors. We decided to keep all the newcomer writers that satisfied our thresholds in order for it to be useful to the newcomer researchers. In both figures the thicker links describe a collaboration that has already happened. The gray links define the proposed associations. We use

TAB. 3: Couples of authors discovered with the hypergraph mining approach, sorted by their corresponding S_{CT} value. In bold, those pairs that have already collaborated at EGC.

Mined latent relations	$\mathbf{S_{CT}}$
Djamel Abdelkader Zighed, Marc Plantevit	1.000
Djamel Abdelkader Zighed, Gilbert Ritschard	0.953
Frédéric Flouvat, Marc Plantevit	0.950
Jean-François Boulicaut, Marc Plantevit	0.950
Gilbert Ritschard, Frédéric Flouvat	0.920
Jean-François Boulicaut, Gilbert Ritschard	0.915
Djamel Abdelkader Zighed, Sandra Bringay	0.865
Sandra Bringay, Frédéric Flouvat	0.785
Jean-François Boulicaut, Sandra Bringay	0.760

TAB. 4: Top 5 couples of topics, represented by their three most prominent concepts sorted by their corresponding S_{CT} value.

Topic-topic relations mined		$\mathbf{S_{CT}}$
T1: ontologie, alignement, sémantique	T9: image, afc, segmentation	0.990
T3: séquence, temporel, événement	T5: document, xml, annotation	0.844
T5: document, xml, annotation	T13: carte, topologique, auto, organisatrice	0.783
T5: document, xml, annotation	T11: donnée, flux, base	0.770
T7: connaissance, gestion, expert	T9: image, afc, segmentation	0.768

$\gamma = 5$ and $\lambda = 7$ for this experiment. In both collaboration graphs, clear sub-structures can be appreciated. They may provide further insights with more experimentation. The author-topic recommendations are omitted due to space constraints. Nonetheless, they will be included in the Web interface of our system, which we describe next.

4 Web interface for exploring the EGC anthology

The interface (available at http://mediamining.univ-lyon2.fr/people/guille/egc2016/) offers 3 ways to have an overview of the articles with: the author index, the complete vocabulary

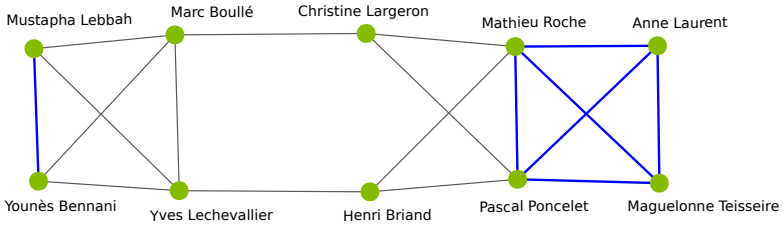

FIG. 9: Recommendation graph for the veteran authors.

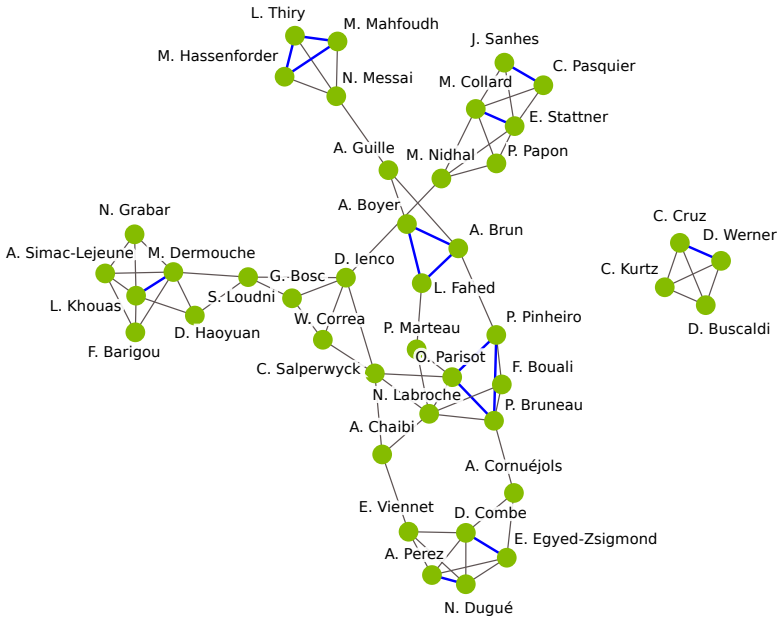

FIG. 10: Recommendation graph for the newcomer authors.

and the topic cloud, where each topic is represented by a bubble labeled with the most relevant words and which diameter is proportional to its overall frequency. It also offers detailed views for: each topic (see Fig. 11), each article, each author (see Fig. 12), and each word of the vocabulary. The detailed view for a topic presents the weight distribution over the most relevant words, the evolution of its frequency through the years, the list of related articles and the collaboration network. The detailed view for an article presents the most significant keywords, the most similar documents, the weight distribution over topics and a link to access the electronic version of the article via the publisher's website. The detailed view about an author displays his/her topic distribution, his/her publication list, as well as personalized collaboration and topic suggestions.

References

Arun, R., V. Suresh, C. V. Madhavan, and M. N. Murthy (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *PAKDD*, pp. 391–402.

Berry, M. W. and M. Browne (2005). Email surveillance using non-negative matrix factorization. *Journal of Computational and Mathematical Organization Theory 11*(3), 249–264.

Blei, D., A. Ng, and M. Jordan (2003). Latent Dirichlet Allocation. *Journal of Machine Learning 3*, 993–1022.

Cavnar, W. and J. Trenkle (1994). N-gram-based text categorization. In *SDAIR*, pp. 161–175.

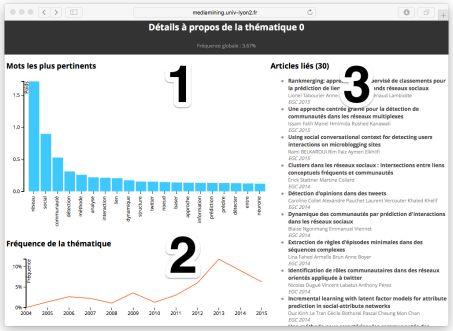

FIG. 11: Details about a topic: (1) most representative words, (2) frequency through time, (3) related articles.

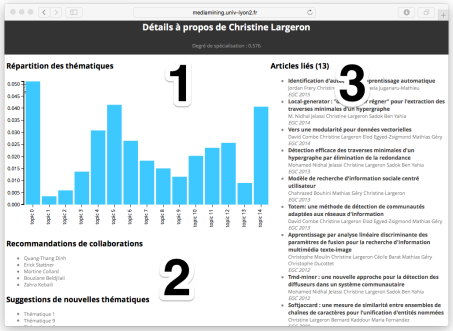

FIG. 12: Details about an author: (1) weight distribution over topics, (2) collaboration and topic suggestions, (3) related articles.

Denis, P. and B. Sagot (2012). Coupling an annotated corpus and a lexicon for state-of-the-art pos tagging. *Language Resources and Evaluation 46*(4), 721–736.

Fouss, F., A. Pirotte, J.-M. Renders, and M. Saerens (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering. 19*(3), 355–369.

Greene, D., D. O'Callaghan, and P. Cunningham (2014). How many topics? stability analysis for topic models. In *ECML PKDD*, pp. 498–513.

Liu, H., P. LePendu, R. Jin, and D. Dou (2011). A hypergraph-based method for discovering semantically associated itemsets. In *ICDM*, pp. 398–406.

Ricci, F., L. Rokach, B. Shapira, and P. B. Kantor (Eds.) (2011). *Recommender Systems Handbook*. Springer.

Sagot, B. (2010). The lefff, a freely available and large-coverage morphological and syntactic lexicon for french. In *LREC*.

Résumé

Dans le cadre du défi proposé à l'édition 2016 de la conférence EGC, nous exploitons les articles qui y ont été publiés de 2004 à 2015, avec pour but d'expliquer sa structure et son évolution. A partir des thématiques latentes découvertes et d'autres propriétés des articles (e.g. auteurs, affiliations), nous mettons en lumière des caractéristiques intéressantes des structures thématique et collaborative d'EGC. A l'aide d'une méthode d'extraction d'itemsets dans les hyper-graphes nous mettons aussi en avant des liens latents entre auteurs ou entre thématiques. De plus, nous proposons des recommandations d'auteurs ou de thématiques. Enfin, nous décrivons une interface Web pour explorer les connaissances découvertes.