University of Jean Monnet

Machine Learning And Data Mining

(MLDM)

EGC Dataset Analysis

Authors:
Mohammad Poul Doust
Robin Khatri
Samaneh Zareian Jahromi

January 15, 2020

Contents

1 Introduction **2**

2 Data Description **2**

3 Objective and studies made **3**
3.1 Related Work 3

4 Exploratory Analysis: "How many authors we have in the dataset ?" **3**
4.1 Objective: 3
4.2 Method: 4
4.3 Results: 4
4.4 Conclusion: 4

5 Exploratory Analysis: "Distribution of articles over time ?" **5**
5.1 Objective: 5
5.2 Method: 5
5.3 Results: 5
5.4 Conclusion: 6

6 "Is there a correlation between authors ?" **6**
6.1 Objective: 6
6.2 Method: 6
6.3 Results: 7
6.4 Conclusion: 10

7 "Hierarchical Clustering on Author-Correlation Matrix" **10**
7.1 Objective: 10
7.2 Method: 10
7.3 Results: 10
7.4 Conclusion: 10

8 "What are the main topics ?" **10**
8.1 Objective: 10
8.2 Method: 10
8.3 Results: 11
8.4 Conclusion: 13

9 "Identify frequent groups of authors ?" 13
9.1 Objective: 13
9.2 Method: 13
9.3 Results: 13
9.4 Conclusion: 13

10 "Identify top authors in case of publishing more articles" 13
10.1 Objective: 13
10.2 Method: 13
10.3 Results: 14
10.4 Conclusion: 14

Abstract

1 Introduction

In this article, we describe work done in the context of scientific analysis of dataset represented in the EGC conference for 20 years , which ultimate goal is to use various data mining and machine learning techniques on analysis of this large dataset. To this purpose, we address some questions as how many authors we have in the dataset, distribution of articles over time, if there is a correlation between authors, hierarchical Clustering on Author-Correlation Matrix, what the main topics are, identifying frequent groups of authors, and identifying top authors in case of publishing more article.

2 Data Description

The dataset analyzed in this report represents articles published in EGC conference from 2004 towards 2018. The dataset follows a tab-separated format. Each row in this file corresponds to an article represented by the following attributes: series, book title of the proceedings, year, title of the article, abstract, authors, URL of a pdf of the first page of the paper, and a URL of the pdf of the full paper.

Each line of the file is a record and the attributes are separated by horizontal tab characters. The data contains some errors (especially in the abstract field): missing spaces, added hyphens, etc. (Probably coming from

Column	Type	Description
Series	Text	conference series for the article
Booktitle	Text	EGC for all rows
Year	Numeric	publish year
Title	Text	title of the article
Abstract	Text	abstract section of the article
Authors	Text	names of the authors
Pdf1page	Text	hyperlink for first page of the article
Pdfarticle	Text	hyperlink for the complete article

Table 1: Dataset Description

the automatic extraction of the abstract from the pdf of the paper). So we preprocessed the data removing stopwords using the entire words in nltk dictionary and doing stemming and lemmatization after that.

3 Objective and studies made

3.1 Related Work

Related research[1] includes investigation of inferring latent topics that pervade this corpus using non-negative matrix factorization, discovering existent but latent relations between authors or between topics through hypergraph itemset extraction process, and proposing topic-author and author- author recommendations with a content-based approach.

4 Exploratory Analysis: "How many authors we have in the dataset ?"

4.1 Objective:

The goal of this section to explore some information about the number of authors in the dataset. Specifically, we see how many unique authors in all

data years and the distribution of authors per year (from 2004 towards 2018).

4.2 Method:

The unique authors are first extracted from the "author" column afterwards a simple data preprocessing is done to split the authors (since the authors are comma delimited). Finally, we process the obtained list of authors to throw duplicates.

4.3 Results:

According the previously mentioned processing, we have 2007 authors in all dataset. They are distributed as depicted in Figure 10. On average, there is around 216 unique authors per year.

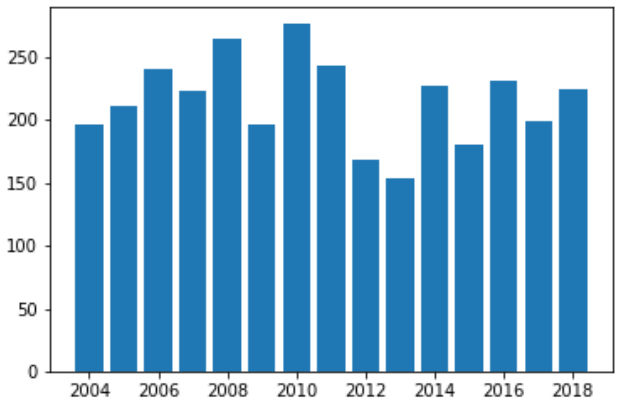

Figure 1: Distribution of authors per year

4.4 Conclusion:

We can see that the number of authors reach its max in year 2010 (more than 250 authors). While its minimum is at year 2013 (around 150 authors). To get more insight about this results, we will explore the distribution of articles per year in the next section.

5 Exploratory Analysis: "Distribution of articles over time ?"

5.1 Objective:

This section is a continuation for the exploratory data analysis done in the previous section. We will check if the distribution of articles over years follows same trend for the number of authors distribution.

5.2 Method:

This is done by simple data processing. We group the dataset on "year" column and we use the COUNT as an aggregation function on the grouped data.

5.3 Results:

We can see from Figure 2. that most of the mass centered in year 2010 (more than 100 article). On the other hand, the least is in year 2013 (less than 60 articles). Moreover, on average there is about 85 articles published on ECG per year.

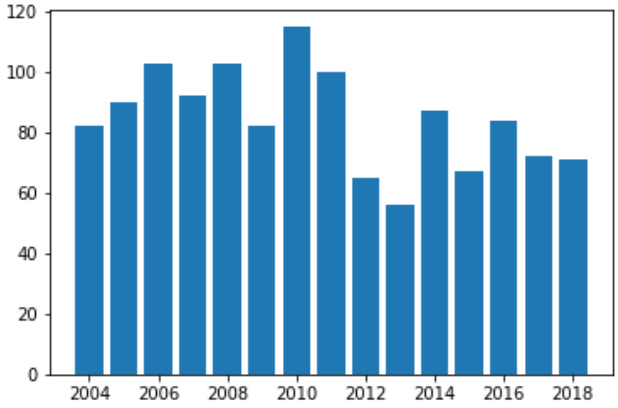

Figure 2: Distribution of articles per year

Title	Year	Abstract	Author_1	Author_2	Author_3	Author_4	Author_5	....
A two level co-clustering ...	2018	.....	False	True	False	True	True	....
Analyse des sentiments...	2018	.....	True	False	False	False	False	....

Table 2: Dataset Transformed

5.4 Conclusion:

We can Conclude that the distribution of articles over years follow the same trends observed in section 4 for the distribution of authors per year. Therefore, there is no trend where we have large number of authors contributing to the same article.

6 "Is there a correlation between authors ?"

6.1 Objective:

The objective of this section is to study the correlation behavior between authors. This correlations is important to see if there is some authors that write together more frequently.

6.2 Method:

In order to find the pair-wise correlation between authors, we need first to do a preprocessing to convert the "author" column into a one-hot encoding format. To that goal, we add to the dataset new column for each author, this column would be a Boolean field corresponding if the some author contribute to a certain article. see Table 2

After transforming the dataset, it easy to study the pairwise correlation between columns by calculating the Pearson Coefficient for the authors columns. At the end, we exclude the self-correlations pairs and get the top-correlated pairs. we can apply this method iteratively by merging each correlated authors and re-calculate the correlations with the others. for this section we will only explore first-level correlations since we will exploit the authors levels through other methods (like clustering). We calculate the Pearson Coefficient according to the equation:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

Where:
cov is the covariance
σ_X is the standard deviation of X
σ_Y is the standard deviation of Y

6.3 Results:

We can see from Figure 3 the heat-map of the author-wise correlation. Since we would have a correlation of 1 for all authors who have only 1 article in all dataset, we tested different thresholds. We can see we have a complete correlation between authors up to 8 articles, those represents authors who have 8 articles in total and all of them are together, which is interesting. Moreover, in all cases, we can see the diagonal red correspond to the self-correlation trend. Additionally, It is worth noticing that we also have some correlation dark-red points (complete correlation factor) aligning vertically or horizontally, those corresponds to the authors correlating to the same author forming a group of correlated authors. Moreover, Table 3 and Figure 4 show a sample of the pair-wise correlation (preserving authors that have exactly 8 articles) after removing self-correlation. It is obvious that we have cross correlation between authors. Finally, Figure 5 depict the number of authors having more than X articles, that is interesting because intuitively, since the data span years from 2004 to 2018, we would expect that each author would have at most one paper in one year, but the figures show that we have around 4 people who has around 20 papers in total, hence, more than one paper in a year.

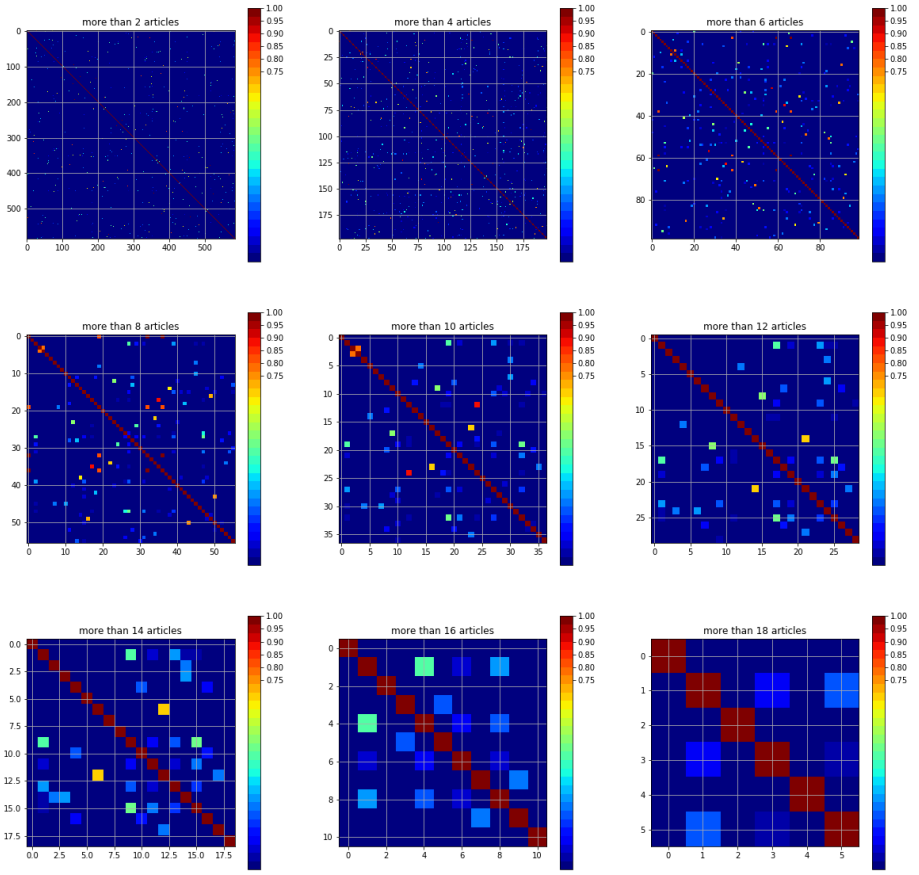

Figure 3: Author-wise correlation with different thresholds (keeping authors that have articles more than a specific threshold)

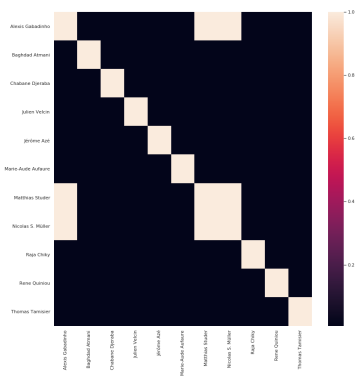

Figure 4: Authors with 8 articles correlation)

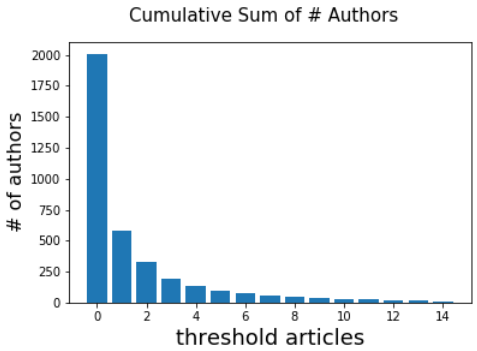

Figure 5: The cumulative sum of authors having different number of articles

Author 1	**Author 2**	**Correlation**
Matthias Studer	Nicolas S. Müller	1.0
Alexis Gabadinho	Matthias Studer	1.0
Alexis Gabadinho	Nicolas S. Müller	1.0
Julien Velcin	Jérôme Azé	0.06

Table 3: Sample of Correlated authors (Having at least 8 articles)

6.4 Conclusion:

We can conclude that the dataset contains authors who are fully-correlated and yet with more than one paper in common (at most 8 papers) which means they always write together. Moreover, there are authors who contributed to more than one paper in each year, which is interesting to find. As this study only checked the pairwise-correlation between authors, it was very clear to observe by visualization the correlation heat-map that we have blobs (indicating high correlation) that are aligned (horizontally or vertically), which represent a group of correlated authors. Therefore, it is worth employing hierarchical clustering algorithm on the correlation matrix to capture those clusters, or we can alternatively apply multi-level correlation.

7 "Hierarchical Clustering on Author-Correlation Matrix"

7.1 Objective:

7.2 Method:

7.3 Results:

7.4 Conclusion:

8 "What are the main topics ?"

8.1 Objective:

The goal of this section is to process the "Abstract" field in order to mine the topics of the articles. To that goal. we will apply the study on all the dataset and using only three years of the dataset separately. To visualize the results, we will be using Wordclouds visualization methods on topic words.

8.2 Method:

The method start by cleaning the abstract fields from the stop words. Since the majority of articles are in French, we use French stop words list from NLTK python library. Afterwords, we tokenize the text and unify cases

(convert to lower case) and perform stemming using Snowball Stemmer[2]. After cleaning the dataset, it is ready for topic modelling. We use Latent dirichlet allocation [3].

LDA is an unsupervised leaning algorithn, Basically, it is a generative probabilistic model approach for collections of discrete data such as text corpora. LDA follows a generative story that consider each document was generated by first choosing a topic distribution for this document. Moreover, each topic has its own word probability distribution.

1. Choose the number of target topics K
2. For each word w in the corpus, draw a topic from Multinomial distribution conditioned on topic
3. to optimize the assignment:
    - For each word w in d:
    - calculate probability p(topic t |document d): the probability distribution of words in document d that are assigned to topic t .
    - Compute p(word w |topic t): the probability distribution of topic assignments over documents d, from word w
4. Reassign word w a new topic t' where we choose t' according to the probability p(topic t' |document d) * P(word w |topic t')

The main goal is to balance two factors:

- For each document d, minimize the number of topics assigned to its words
- For each topic t, minimize the number of terms assigned to t

8.3 Results:

Figures 6, 7, 8, 9 shows the Wordclouds for different cases in the dataset. the Wordclouds were built using the topic words extracted from LDA algorithm by choosing on average 10 topics for each case.

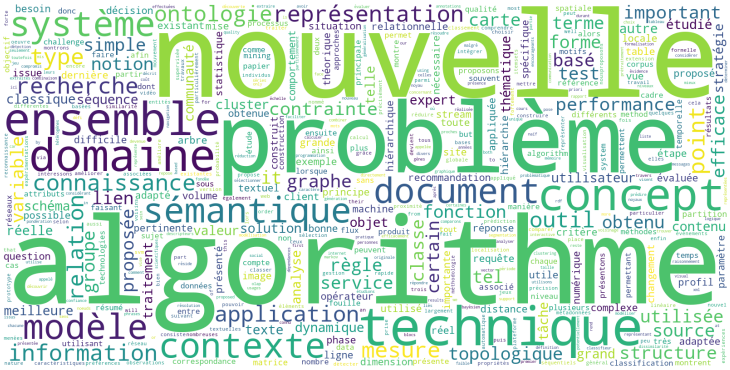

Figure 6: Wordcloud for topics in all years

Figure 7: Wordcloud in 2010

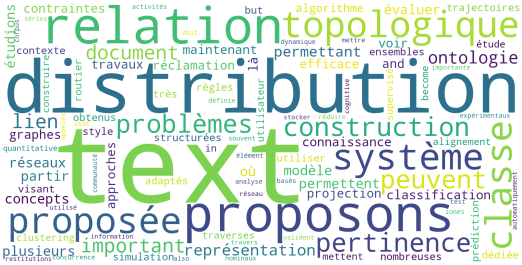

Figure 8: Wordcloud in 2013

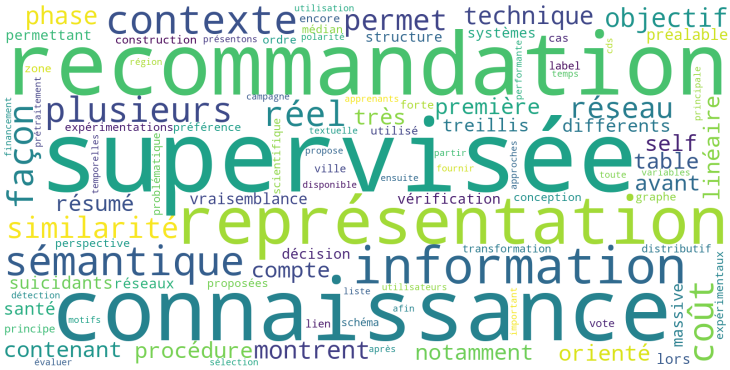

Figure 9: Wordcloud for topics in 2018

8.4 Conclusion:

The Wordcloud visualization provides a neat way to get an abstract idea about the text data. Moreover, when combined with topic modelling it will give more refined idea. For example we can conclude from the figures above that the dataset in general is about text mining, also in 2010m expert system seems to be the trend while in 2010 it's biased towards conceptual relations and ontologies. Finally, in 2018, recommendation systems and supervised machine learning seems to be dominating on the topics

9 "Identify frequent groups of authors ?"

9.1 Objective:

9.2 Method:

9.3 Results:

9.4 Conclusion:

---

10 "Identify top authors in case of publishing more articles"

10.1 Objective:

The goal of this section is to explore some information about the authors in the dataset who published more articles.

10.2 Method:

At first, we preprocessed the author filed to distinguish every set of authors participated in writing an article. After that through a mapper function , we extracted the number of times every author published. Finally, in a reducer function we sorted the obtained list of authors to find the top authors.

10.3 Results:

According the previously mentioned processing, the top authors are listed in figure 10.

```
{'claudia marinica': 2,
 'julien longhi': 1,
 'nader hassine': 1,
 'abdulhafiz alkhouli': 1,
 'boris borzic': 1,
 'marius barctus': 1,
 'marc boullé': 36,
 'fabrice clérot': 16,
 'jean-benoît griesner': 3,
 'talel abdesssalem': 4,
 'hubert naacke': 4,
 'pierre dosne': 1,
 'abdeljalil elouardighi': 1,
 'mohcine maghfour': 1,
 'hafdalla hammia': 1,
 'fatima-zahra aazi': 3,
 'elyase lassouli': 1,
```

(a) Mapper output

```
[('marc boullé', 36),
 ('pascal poncelet', 25),
 ('maguelonne teisseire', 23),
 ('mustapha lebbah', 21),
 ('christine largeron', 18),
 ('mathieu roche', 18),
 ('amedeo napoli', 17),
 ('yves lechevallier', 17),
 ('anne laurent', 17),
 ('fabrice clérot', 16),
 ('younès bennani', 16)]
```

(b) Reducer output

Figure 10: The number of published articles per author through map-reduce functions

10.4 Conclusion:

We can see that the author who published the most than others is Marc Boullé with 36 articles.

References

[1] Ciprian-Octavian Truica Adrien Guille, Edmundo-Pavel Soriano-Morales. Topic modeling and hypergraph mining to analyze the egc conference history, 2015.

[2] Martin F Porter. Snowball: A language for stemming algorithms, 2001.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.