Topic Modelling in Social Networks with Formal Concept Analysis

CMMSE 2021

P. Cordero[1], M. Enciso[2], **D. López-Rodríguez**[1], Á. Mora[1]

[1]Dep. de Matemática Aplicada
[2]Dep. de Lenguajes y Ciencias de la Computación

UNIVERSIDAD DE MÁLAGA

Table of Contents

Introduction

Objectives of this work:

- To present an approach to perform topic modelling on text documents by means of FCA
- To present the idea of clustering the concepts of a formal context, with a new *semidistance*

Topic modelling

In text mining, *topic modelling* is the process of *clustering* on collections of documents, that is, dividing a series of documents into groups or *clusters* that may arise *naturally*, so that each cluster represents a topic that can be understood or studied independently of the others.

It is therefore a method of unsupervised classification of text documents.

Main techniques

Topic modelling is based on two principles:

- Each document is a mixture of several topics. For example, considering a model to find 2 topics in a collection of documents, we could say "document 1 is 90% topic A and 10% topic B, while document 2 is 30% topic A and 70% topic B".
- Each topic is a mixture of words, which are in different proportions in the different topics.

Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) are mathematical methods for the joint estimation of these proportions: adjusting the *mix* of words that are associated with a topic, while determining for each document its belonging to each of the topics.

Formal Concept Analysis

- Strongly based on logic and lattice theory
- Allows to represent the knowledge inside a *dataset* in two ways:
  - Implications, that is, exact association rules between the attributes (variables) of a problem
  - Concepts: the closed sets of attributes ($\equiv$ the closed itemsets in transactional databases).

Document Analysis with FCA

We use as dataset the binary Document-Term matrix (DTM):

- The rows (objects) are documents
- The columns (attributes) are the terms
- A one in a cell of the table indicates the presence of the given term in the corresponding document.

Instead of estimating the topics of a set of documents, we will extract the lattice of concepts that model the internal knowledge of the documents from this DTM, and perform topic modelling on the concepts themselves.
This means that we will perform *clustering* of the set of concepts, and transfer this *clustering* to the documents.

Clustering concepts. Why?

- The concept lattice represent all the possible closed sets of terms.
- A concept is formed by terms that appear jointly in *all* the documents.
- A closed set (concept) of terms is shown to be a maximal rectangular bicluster of terms and documents.
- Thus, the clustering of concepts takes into account more information/knowledge than the simple clustering of documents.

The procedure

As in classical clustering, we need to define a *semidistance* between the concepts:

$$d(C_1, C_2) = \sigma(C_1 \vee C_2 : C_1) + \sigma(C_1 \vee C_2 : C_2)$$

where $\vee$ is the *supremum* operator (of two concepts, in the concept lattice), and $\sigma(A : B)$ is the operator that gives the "minimum chain length between A and B".
To compute $d(C_1, C_2)$:

- We take the supremum of C_1 and C_2, we call it C.
- We compute the shortest chains from C to C_1 and C_2.
- The sum of their lengths is $d(C_1, C_2)$.

Once the distance is computed between all pairs of concepts, the clustering can be performed by running any appropriate algorithm.
In this work, we use the *Partition Around Medoids* algorithm, which, iteratively:

- Find the optimal partition into k clusters
- Find the central item in each cluster, that is, it will provide the central concepts related to document topics.

To cluster the documents back:

- Each document is an object
- We compute its *object concept*
- We assign the document to the cluster given by its associated *object concept*.

Real example

- More than 9000 tweets from two hashtags: “#covid” and “#trump”, after the 2020 USA presidential elections.
- From these tweets, once the tweets with few terms were eliminated (1658 remaining), the most frequent terms (those that appear in at least 0.5% of the tweets) were extracted (80 terms or attributes), and the DTM was constructed.
- We compute the concept lattice, with 3510 concepts, and compute all the pairwise distances.
- Then, we cluster these concepts using PAM.

The terms in the central concepts were:

1. `covid`
2. `trump`

Since the clustering was unsupervised, the procedure has been able to detect automatically the topics.
Now, we want to relate the clusters to the original hashtags, to confirm the relationship.

- Accuracy: measures the correct assignment of cluster to hashtag.
- Purity: measures how terms are assigned to one topic
- Coherence: measures that terms that co-appear often are assigned to the same topic
- Contrast: measures how topics are strongly related to the terms it contains.

Results

Method	Accuracy	Purity	Coherence	Contrast
LDA	0.8661	3.4898	0.367	0.5979
LSA	0.9240	3.5532	0.223	0.5416
Proposal	0.9910	3.5465	0.367	0.5425

Conclusions and future work

- FCA can be used with document-term matrices as formal contexts in order to obtain valuable knowledge from collections of documents.
- We have presented the idea of concept clustering.
- This idea is based on the use of a novel *semidistance* on the concept lattice.
- The *semidistance* together with an appropriate clustering algorithms allow us to group similar concepts.
- This can be extrapolated to topic modelling in social network analysis, with promising results, at least comparable to the current state-of-the-art.
- As future work, we will study the relationship of the defined *semidistance* to other metrics that could be defined on the concept lattice, and to analyse the use of graph theory to gain a better understanding of the concept lattice.

Topic Modelling in Social Networks with Formal Concept Analysis

CMMSE 2021

P. Cordero[1], M. Enciso[2], **D. López-Rodríguez**[1], Á. Mora[1]

[1]Dep. de Matemática Aplicada
[2]Dep. de Lenguajes y Ciencias de la Computación

UNIVERSIDAD
DE MÁLAGA