A multi-criteria document clustering method based on topic modeling and pseudoclosure function

Quang Vu Bui
Hue University of Sciences
CHArt Laboratory EA 4004
EPHE, Paris, France
quang-vu.bui@etu.ephe.fr

Karim Sayadi
Sorbonne University, UPMC
Univ Paris 06
CHArt Laboratory EA 4004
EPHE, Paris, France
karim.sayadi@upmc.fr

Marc Bui
EPHE and UP8 Univ Paris 08
CHArt Laboratory EA 4004
Paris, France
marc.bui@ephe.sorbonne.fr

ABSTRACT

We address in this work the problem of document clustering. Our approach is based on the following pipeline. First, we quantify the topics in a document. Then, a number of clusters is set automatically. Finally, a multi-criteria distance is defined to cluster the documents. The advantage of this approach is that it allows us to have a number of multi-criteria clusters based on structural analysis of each document. We have applied our method on Twitter data and showed the accuracy of our results compared to a random choice number of clusters.

CCS Concepts

•**Information systems** → **Data mining; Clustering; Information retrieval;** *Document topic models;*

Keywords

Latent Dirichlet Allocation, Topic modeling, Gibbs Sampling, pretopology, pseudoclosure, clustering, k-means.

1. INTRODUCTION

Classifying a set of documents is a standard problem addressed in machine learning and statistical natural language processing [8]. The classical approach for classifying documents consists in the use of a measure of similarity. Depending on the algorithm, different measures are used. In this work, we tackle the problem of the classification of documents in a different way by defining a family of binary relationships on the topic-based contents of the documents. The documents are not only classified using a measure of similarity but also using pseudoclosure function built from family of binary relationships between the different hidden semantic contents (i.e topics) computed by the Latent Dirichlet Allocation (LDA). The pseudoclosure function are defined using an original concept called pretopology [1].

SoICT 2015, December 03-04, 2015, Hue City, Viet Nam
 ISBN 978-1-4503-3843-1/15/12...$15.00
DOI: http://dx.doi.org/10.1145/2833258.2833291

LDA is a generative probabilistic model by Blei et al. [2] that allows us to discover the latent structure (i.e. the topic structure) of a set of documents. LDA gives us three latent variables after computing the posterior distribution of the model; the topic assignment, the distribution of words in each topic and the distribution of the topics in each document. Having the distribution of topics in documents, the pretopology allows to compute the elements of each cluster of documents using the pseudoclosure function which gives us the ability to follow the process step by step. The pseudoclosure function is applied on the distribution of the topic, computed by LDA, in each document. We use the pretopology for its ability to manipulate many binary relationships. This framework allows us first, to model a topic space and to cluster the documents according to their positions in the topic space.

We present this connection by a method that we named the Method of Clustering Documents using Pretopology and Topic Modeling (MCPTM). MCPTM organizes a set of unstructured entities in a number of clusters based on multiple relationships between each two entities. Our method discovers the topics expressed by the documents, tracks changes step by step over time, expresses similarity based on multiple criteria and provides both quantitative and qualitative measures for the analysis of the document.

The continuation of this article is organized as follows: section 2, 3 present some basic concepts such as Latent Dirichlet Allocation (section 2) and the Pretopology theory (section 3), section 4 explains our approach by describing at a high level the different parts of our algorithm. In the section 5, we apply our algorithm to a corpus consisting of microblogging posts that comes from Twitter.com. We conclude our work in section 6 by presenting the obtained results.

2. TOPIC MODELING

Topic modeling is a method for analyzing large quantities of unlabeled data. For our purposes, a topic is a probability distribution over a collection of words and a topic model is a formal statistical relationship between a group of observed and latent (unknown) random variables that specifies a probabilistic procedure to generate the topics [2, 5, 3, 11]. In many cases, there exists a semantic relationship between terms that have high probability within the same topic – a phenomenon that is rooted in the word co-occurrence patterns in the text and that can be used for information retrieval and knowledge discovery in databases.

2.1 The Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) by Blei et al. [2] is a generative probabilistic model for collections of grouped discrete data. Each group is described as a random mixture over a set of latent topics where each topic is a discrete distribution over the collections's vocabulary. LDA is applicable to any corpus of grouped discrete data, we will refer to the standard Natural Language Processing (NLP) use case where a corpus is a collection of documents, and the data are words.

LDA is a probabilistic model for unsupervised learning, it can be seen as a Bayesian extension of the probabilistic Latent Semantic Analysis (pLSA) [5]. More precisely, LDA defines a complete generative model which with a uniform prior is a full bayesian estimator while pLSA provides an Maximum Likelihood (ML) or Maximum a Posterior (MAP) estimator. For more technical details refer to the work of Gregor Heinrich [4]. The generative model of LDA is described with the probabilistic graphical model [6] in Fig. 1.

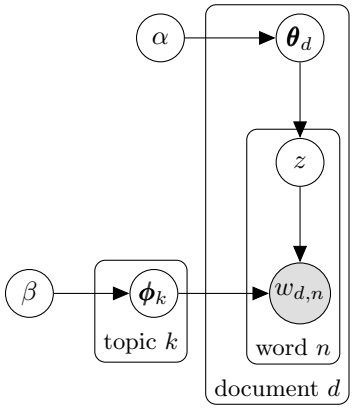

Figure 1: Bayesian Network (BN) of Latent Dirichlet Allocation.

In LDA model, different documents d have different topic proportions θ_d. In each position in the document, a topic z is then selected from the topic proportion θ_d. Finally, a word is picked from all vocabularies based on their probabilities ϕ_k in that topic z.

The advantage of the LDA model is that examining at the topic level instead of the word level allows us to gain more insights into the meaningful structure of documents, since noise can be suppressed by the clustering process of words into topics. Consequently, we can utilize the topic proportion in order to organize, search, and classify a collection of documents more effectively.

2.2 Inference with Gibbs sampling

In this subsection, we specify a topic model procedure based on the Latent Dirichlet Allocaiton (LDA) and the Gibbs Sampling.

The key problem in topic modeling is posterior inference. This refers to reversing the defined generative process and learning the posterior distributions of the latent variables in the model given the observed data. In LDA, this amounts to solve the following equation:

$$p(\theta, \phi, z|w, \alpha, \beta) = \frac{p(\theta, \phi, z, w|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{1}$$

Unfortunately, this distribution is intractable to compute [4]. The normalization factor in particular, $p(w|\alpha, \beta)$, cannot be computed exactly. However, as there are a number of approximate inference techniques available that we can apply to the problem including variational inference (as used in the original LDA paper [2]) and Gibbs Sampling that we propose to use.

For LDA, we are interested in the latent document-topic proportions θ_d, the topic-word distributions $\phi^{(z)}$, and the topic index assignments for each word z_i. While conditional distributions - and therefore an LDA Gibbs Sampling algorithm - can be derived for each of these latent variables, we note that both θ_d and $\phi^{(z)}$ can be calculated using just the topic index assignments z_i (i.e. z is a sufficient statistic for both these distributions). Therefore, a simpler algorithm can be used if we integrate out the multinomial parameters and simply sample z_i. This is called a collapsed Gibbs sampler [3, 11].

The collapsed Gibbs sampler for LDA needs to compute the probability of a topic z being assigned to a word w_i, given all other topic assignments to all other words. Somewhat more formally, we are interested in computing the following posterior up to a constant:

$$p(z_i \mid z_{-i}, \alpha, \beta, w) \tag{2}$$

where z_{-i} means all topic allocations except for z_i.

Require: words $w \in$ corpus $\mathcal{D} = (d_1, d_2, \ldots, d_m)$
procedure LDA-GIBBS(w, α, β, T)
    randomly initialize z and increment counters
    loop for each iteration
        loop for each word w in corpus $\mathcal{D}$
            Begin
                word $\leftarrow w[i]$
                tp $\leftarrow z[i]$
                $n_{d,tp} -= 1; n_{word,tp} -= 1; n_{tp} -= 1$
                loop for each topic j $\in \{0, \ldots, K-1\}$
                    compute $P(z_i = j|z_{-i}, w)$
                $tp \leftarrow sample \quad from \quad p(z|.)$
                $z[i] \leftarrow tp$
                $n_{d,tp} += 1; n_{word,tp} += 1; n_{tp} += 1$
            End
    Compute $\phi^{(z)}$
    Compute θ_d
    return $z, \phi^{(z)}, \theta_{\mathcal{D}}$ ▷ Output
end procedure

Figure 2: Algorithm 1: The LDA Gibbs sampling algorithm.

Equation 3 computation of the posterior distribution for topic assignment.

$$P(z_i = j|z_{-i}, w) \propto \frac{n^{w_i}_{-i,j} + \beta}{n^{(\cdot)}_{-i,j} + V\beta} \frac{n^{d_i}_{-i,j} + \alpha}{n^{d_i}_{-i,\cdot} + K\alpha} \tag{3}$$

where $n^{w_i}_{-i,j}$ is the number of times word w_i was related to topic j. $n^{(\cdot)}_{-i,j}$ is the number of times all other words were related with topic j. $n^{d_i}_{-i,j}$ is the number of times topic j was related with document d_i. And $n^{d_i}_{-i,\cdot}$ is the number of times all other topics were related with document d_i. Those notations were taken from the work of Thomas Griffiths and Mark Steyvers [3].

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + V\beta} \quad (4)$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + K\alpha} \quad (5)$$

Equation 4 is the bayesian estimation of the distribution of the words in a topic. Equation 5 is the estimation of the distribution of topics in a document.

3. PRETOPOLOGY THEORY

The pretopology is a mathematical modeling tool for the concept of proximity in the field of social sciences in the discrete spaces [1]. It probably establishes the powerful tools for the structure analysis and automatic classification. It ensures the follow-up of the process development of dilation, alliance, adherence, closed subset, acceptability [12, 7].

3.1 Pseudoclosure

DEFINITION 1. *We call pseudoclosure defined on E, any function $a(.)$ from $\mathcal{P}(E)$ into $\mathcal{P}(E)$ such as:*

$$a(\emptyset) = \emptyset; \forall A \subset E, A \subset a(A) \quad (6)$$

Then, (E, a) is said a pretopology space.

According to properties of $a(.)$, we obtain more or less complex pretopological spaces from the most general spaces to topological spaces. Pretopological spaces of $\mathcal{V}$- type are the most interesting cases.

DEFINITION 2. *A pretopology space (E, a) is called $\mathcal{V}$-type space if and only if*

$$\forall A \subset E, \forall B \subset E, (A \subset B) \Rightarrow (a(A) \subset a(B)) \quad (7)$$

3.2 Pretopology and binary relationships

Suppose we have a family $(R_i)_{i=1,\ldots,n}$ of binary reflexive relationships on a finite set E. Let us consider $\forall i = 1, 2, \ldots, n, \forall x \in E, V_i(x)$ defined by:

$$V_i(x) = \{y \in E | x \, R_i \, y\} \quad (8)$$

Then, the pseudoclosure $a_s(.)$ is defined by:

$$\mathbf{a_s}(A) = \{x \in E | \forall i = 1, 2, \ldots, n, V_i(x) \cap A \neq \emptyset\} \quad (9)$$

Pretopology defined on E by $a_s(.)$ using the intersection operator is called the strong pretopology induced by the family $(R_i)_{i=1,\ldots,n}$.

Similarly, we can define weak pretopology from $a_w(.)$ by using union operator:

$$\mathbf{a_w}(A) = \{x \in E | \exists i = 1, 2, \ldots, n, V_i(x) \cap A \neq \emptyset\} \quad (10)$$

PROPOSITION 1. *$a_s(.)$, $a_w(.)$ determine on E a pretopological structures and the spaces (E, a_s), (E, a_w) are of $\mathcal{V}$-type.*

3.3 Minimal closed subsets

DEFINITION 3. *Let (E, a) a pretopological space, $\forall A, A \subset E$. A is a closed subset if and only if $a(A) = A$.*

DEFINITION 4. *Given (E, a) a pretopological space, call the closure of A, when exists, the smallest closed subset of (E, a) which contains A. The closure of A is denoted by $F(A)$.*

PROPOSITION 2. *In any pretopological space of type $\mathcal{V}$, given a subset A of E, the closure of A always exists.*

So, in the $\mathcal{V}$-type space, given a set finite E, the closure F(A) is always exits and can be calculated by using the following property that is useful in calculating distance between elements.

$$\exists k < |E|, F(A) = a^k(A) = a(a^{k-1}(A))$$

We denote $\mathcal{F}_e$ is the family of elementary closed subsets, the set of closures of each singleton $\{x\}$ of $P(E)$. So in a $\mathcal{V}$-type pretopological space, we get:

- $\forall x \in E, \exists F_x$: closure of $\{x\}$.
- $\mathcal{F}_e = \{F_x | x \in E\}$

DEFINITION 5. *F is called a minimal closed subset if and only if F is a minimal element for inclusion in $\mathcal{F}_e$.*

We denote $\mathcal{F}_m = \{F_{m_j}, j = 1, 2, \ldots, k\}$, the family of minimal closed subset, the set of minimal closed subsets in $\mathcal{F}_e$.

4. OUR APPROACH

In our approach, we build The Method of Clustering Documents using Pretopology and Topic Modeling (MCPTM) which clusters documents via the topic modeling and pseudoclosure. MCPTP can be built in following:

1. Defining topic-distribution of each document d_i in corpus $\mathcal{D}$ by document structure analysis using LDA.
2. Defining two binary relationships: R_{MTP} based on major topic and R_{d_H} based on Hellinger distance.
3. Building pseudoclosure function from two binary relationships R_{MTP}, R_{d_H}.
4. Building pseudoclosure distance from pseudoclosure function.
5. Determining initial parameters for k-means algorithm from results of minimal closed subsets.
6. Using *k-means* algorithm to cluster set of documents with initial parameters from result of minimal closed subsets, pseudoclosure distance to compute the distance between two objects and inter-pseudoclosure distance to re-compute the new centroids.

4.1 Document structure analysis by LDA

A term-document matrix is given as an input to LDA and it outputs two matrices, the document-topic distribution matrix θ and the topic-term distribution matrix ϕ. The topic-term distribution matrix $\phi \in \mathbf{R}^{KxV}$ consists of K rows, where the i-th row $\phi \in \mathbf{R}^V$ is the word distribution of topic i. The terms with high ϕ_{ij} values indicate that they are the representative terms of topic i. Therefore, by looking at such terms one can grasp the meaning of each topic without looking at the individual documents in the cluster.

In a similar way, the documents-topics distributions matrix $\theta \in \mathbf{R}^{MxK}$ consists of M rows, where the i-th row $\theta_i \in \mathbf{R}^K$ is the topic distribution for document i. A high probability value of θ_{ij} indicates that document i is closely related to topic j. In addition, documents with low θ_{ij} values over all the topics are noisy documents that belong to none of the topics. Therefore, by looking at the θ_{ij} values, one can understand how closely the document is related to the topic. We have often been interested in the major topic of document. So, we define the major-topic of each document such as:

DEFINITION 6. *We call $MTP(d_i)$ is major-topic of document d_i if $MTP(d_i)$ is the topic having the largest probability in topic-distribution of document d_i and this probability greater than p_0, $p_0 \geq 1/K$, K is the number of topic.*
$MTP(d_i) = \{k|\theta_{ik} = max_j\theta_{ij} \quad and \quad \theta_{ik} \geq p_0\}$.

Two documents d_m, d_n with their major-topic $MTP(d_m)$, $MTP(d_n)$, are close to each other if they have the same major-topic. A first document clustering can be computed based on the set of major-topics.

4.2 Defining binary relationships

4.2.1 Based on major topic

Considering two documents d_m,d_n with their major-topic $MTP(d_m)$, $MTP(d_n)$, we see that document d_m is "near" to document d_n if they have the same major-topic. So, we proposed a definition of binary relationship R_{MTP} of two documents based on their major-topic such as:

DEFINITION 7. *We call document d_m have binary relationship R_{MTP} with document d_n if d_m and d_n have the same major-topic.*

4.2.2 Based on Hellinger distance

By using LDA, each document is characterized by its topic distribution. The similarity of two documents is measured as the distance between the two corresponding probability distributions. If we consider a probability distribution as a vector, we can choose some distances or similarity measures related to the vector distance such as Euclide distance, Cosine Similarity, Jaccard Coefficient, Pearson Correlation Coefficient, etc. But, it is better if we choose distances or similarity measures related to the probability distribution such as Kullback-Leibler Divergence, Bhattacharyya distance, Hellinger distance, etc. For our work, we choose the Hellinger distance because it is a metric for measuring the deviation between two probability distributions, we can easily compute it and it is especially limited in $[0, 1]$.

DEFINITION 8. *For two discrete probability distributions $P = (p_1, \ldots, p_k)$ and $Q = (q_1, \ldots, q_k)$, their Hellinger distance is defined as*

$$d_H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{k} (\sqrt{p_i} - \sqrt{q_i})^2}, \tag{11}$$

Hellinger distance is directly related to the Euclidean norm of the difference of the square root vectors, i.e.

$$d_H(P,Q) = \frac{1}{\sqrt{2}} \left\| \sqrt{P} - \sqrt{Q} \right\|_2.$$

The Hellinger distance satisfies the inequality of $0 \leq d_H \leq 1$. This distance is a metric for measuring the deviation between two probability distributions. The distance is 0 when $P = Q$. Disjoint P and Q shows the maximum distance of 1. The lower value of Hellinger distance, the smaller deviation between two probability distributions. So, we can use the Hellinger distance to measure the similarity between two documents d_m,d_n. We then define binary relationship R_{d_H} between two documents such as:

DEFINITION 9. *We call document d_m have binary relationship R_{d_H} with document d_n if $d_H(d_m, d_n) \leq d_0, 0 \leq d_0 \leq 1$.*

We can also use the document-topic distribution matrix θ as input and Hellinger distance as distance measure for k-means algorithm to cluster documents.

4.3 Building pseudoclosure function

Based on two binary relationships R_{MTP} and R_{d_H}, we can build neighborhood basis (Algorithm 2, Fig. 3) and then build pseudoclosure (Algorithm 3, Fig. 4) for strong (with intersection operator) and weak (with union operator) Pretopology.

Require: document-topic distribution matrix θ, corpus $\mathcal{D}$
Require: R_{MTP}, R_{d_H}: family of relations.
1: **procedure** NEIGHBORHOOD-TM($\mathcal{D}, \theta, R_{MTP}, R_{d_H}$)
2: **loop** for each relation $R_i \in \{R_{MTP}, R_{d_H}\}$
3: **loop** for each document $d_m \in \mathcal{D}$
4: **loop** for each document $d_n \in \mathcal{D}$
5: **If** $R_i(d_m, d_n)$ **then**
6: $B_i[d_m].append(d_n)$
7: **return** $B = [B_1, B_2]$ ▷ Output
8: **end procedure**

Figure 3: Algorithm 2: Neighborhood Basis Using Topic Modeling.

Require: $B = (B_1, B_2), \mathcal{D} = \{d_1, \ldots, d_m\}$
1: **procedure** PSEUDOCLOSURE($A, B, \mathcal{D}$)
2: aA=A
3: **loop** for each document $d_n \in \mathcal{D}$
4: **If** $(A \cap B_1[d_n] \neq \emptyset \quad or \quad A \cap B_2[d_n] \neq \emptyset)$ **then**
5: $aA.append(d_n)$
6: **return** aA ▷ Ouput
7: **end procedure**

Figure 4: Algorithm 3: Pseudoclosure using Topic Modeling.

4.4 Building pseudoclosure distance

In standard k-means algorithm, the centroid of a cluster is the average point in the multidimensional space. Its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster which is not effective with categorical data analysis. In the other hand, the pseudoclosure distance is used to examine the similarity for both numeric and categorical data. Therefore, it can contribute to improve the classification with k-means.

DEFINITION 10. *We define $\delta(A, B)$ pseudoclosure distance between two subsets A and B of a finite set E:*

$$k_0 = \min(\min\{k|A \subset a^k(B)\}, \infty)$$

$$k_1 = \min(\min\{k|B \subset a^k(A)\}, \infty)$$

$$\delta(A, B) = \min(k_0, k_1)$$

where $\mathbf{a}^k(.) = \mathbf{a}^{k-1}(\mathbf{a}(.))$

DEFINITION 11. *We call $D_A(x)$ interior-pseudo-distance of a point x in a set A:*

$$D_A(x) = \frac{1}{|A|} \sum_{y \in A} \delta(x, y).$$

In case where A and B are reduced to one element x and y, we get the distance $\delta(x, y)$. For clustering documents with k-means algorithm, we use pseudoclosure distance $\delta(x, y)$ to compute distance between two documents (each document represented by its topic-distribution is a point $x \in E$) and interior-pseudo-distance $D_A(x)$ to compute centroid of A (x_0 is chosen as centroid of A if $D_A(x_0) = min_{x \in A} D_A(x)$).

4.5 Structure analysis with minimal closed subsets

The two limits of standard k-means algorithm are the number of clusters which must be predetermined and the randomness in the choice of the initial centroids of the clusters. Pretopology theory gives a good solution to omit these limits by using the result from minimal closed subsets. The algorithm to compute minimal closed subset is presented in Fig. 5, algorithm 4.

Require: corpus $\mathcal{D}$, pseudoclosure $aA()$
procedure MINIMAL-CLOSED-SUBSETS($\mathcal{D}, aA()$)
  compute family of elementary closed subsets $\mathcal{F}_e$
  $\mathcal{F}_m = \emptyset$
  loop until $\mathcal{F}_e = \emptyset$
    Begin
      Choose $F \subset \mathcal{F}_e$
      $\mathcal{F}_e = \mathcal{F}_e - F$
      minimal = True
      $\mathcal{F} = \mathcal{F}_e$
      loop until $\mathcal{F} = \emptyset$ and not minimal
        Begin
          Choose $G \in \mathcal{F}$
          If $G \subset F$ **then**
            minimal=False
          Else
          If $F \subset G$ **then**
            $\mathcal{F}_e = \mathcal{F}_e - \{G\}$
          $\mathcal{F} = \mathcal{F} - G$
        End
    End
  If minimal =True &&$F \notin \mathcal{F}_m$ **then**
    $\mathcal{F}_m = \mathcal{F}_m \cup F$
  return $\mathcal{F}_m$ ▷ Ouput
end procedure

Figure 5: Algorithm 4: Minimal closed subsets algorithm.

By performing the minimal closed subset algorithm, we get the family of minimal closed subsets. This family, by definition, characterizes the structure underlying the data set E. So, the number of minimal closed subsets is a quite important parameter: it gives us the number of clusters to use in the k-means algorithm. Moreover, the initial centroids for starting k-means process can be determined by using interior-pseudo-distance for each minimal closed subset $F_{m_j} \in \mathcal{F}_m$ (x_0 is chosen as centroid of F_{m_j} if $D_{F_{m_j}}(x_0) = min_{x \in F_{m_j}} D_{F_{m_j}}(x)$).

4.6 MCPTM algorithm

In this subsection, we present The Method of Clustering Documents using Pretopology and Topic Modeling (MCPTM) which clusters documents via the topic modeling and pseudoclosure. At first, an LDA topic modeling is learned on the documents to achieve topic-document distributions. The major topic and Hellinger probability distance are used to define relations between documents and these relations are used to define a pretopological space which can be employed to get preliminarily clusters of a corpus and determine the number of clusters. After that, k-means clustering algorithm is used to cluster the documents data with pseudodistance and inter-pseudodistance. The MCPTP algorithm is presented in Fig. 6, algorithm 5.

Require: $\mathcal{D}$: corpus from set of documents
procedure MCPTP($\mathcal{D}$)
  $\theta_{\mathcal{D}} \leftarrow$ LDA-GIBBS($\mathcal{D}$, α, β, T)
  $B \leftarrow$ NEIGHBORHOOD-TM($\mathcal{D}, \theta_{\mathcal{D}}$, R_{MTP}, R_{d_H})
  $aA \leftarrow$ PSEUDOCLOSURE(B)
  $\mathcal{F}_m \leftarrow$ MIMINAL-CLOSED-SUBSETS($\mathcal{D}$, $aA()$)
  $k = |\mathcal{F}_m|$: number of clusters
  $M = \{m_i\}_{,i=1,\ldots,k}, m_i = Centroid(F_{m_i})$
  while clusters centroids changed **do**
    for each x $\in E - M$ **do**
      compute $\delta(x, m_i), i = 1, \ldots, k$
      find m_0 with $\delta(x, m_0) = min\delta(x, m_i)_{i=1,\ldots,k}$
      $F_{m_0} = F_{m_0} \cup \{x\}$
    end for
    Recompute clusters centroids M.
  end while
  return $Clusters = \{F_1, F_2, \ldots, F_k\}$ ▷ Output
end procedure

Figure 6: Algorithm 5: The MCPTM algorithm: clustering documents using pretopology and topic modeling.

4.7 Implementation in python of the library AMEUR

In this part, we briefly present our *AMEUR* library written in python. *AMEUR* is a project connecting the tools that come from the framework of pretopology, topic modeling, multi-relations networks analysis and semantic relationship. The library is composed of the following modules: *pretopology*, *topicmodeling* and *nlp*.

The *pretopology* module implements the functions described in section III. The implementation of the pretopology in the *AMEUR* library allows us to ensures the follow-up of step-by-step processes like dilatation, alliance, pseudoclosure, closure, family of minimal closed subsets, MCPTP and acceptability in multi-relations networks.

The *topicmodeling* module implements generative models like the Latent Dirichlet Allocation, LDA Gibbs Sampling that allows us to capture the relationships between discrete data. This module is used within the *AMEUR* library for querying purposes e.g to retrieve a set of documents that are relevant to a query document or to cluster a set of document given a latent-topic query. These computation of these queries are insured by the connection between the *topicmodeling* module and the *pretopology* module.

The *nlp* (natural language processing) module implements the necessary functions for getting unstructured text data of different sources from webpages or social medias and preparing them as a proper inputs for the algorithms implemented in the rest of the modules of the library.

Table 1: Words - Topic distribution ϕ and the related users from the θ distribution

Topic 3					Topic 10				
Words	Prob.	Users	ID	Prob.	Words	Prob.	Users	ID	Prob.
paris	0.008	GStephanopoulos	42	0.697	ces	0.010	bxchen	22	0.505
charliehebdo	0.006	camanpour	23	0.694	people	0.007	randizuckerberg	102	0.477
interview	0.006	AriMelber	12	0.504	news	0.006	NextTechBlog	88	0.402
charlie	0.005	andersoncooper	7	0.457	media	0.006	lheron	71	0.355
attack	0.005	brianstelter	20	0.397	tech	0.006	LanceUlanoff	68	0.339
warisover	0.004	yokoono	131	0.362	apple	0.006	MarcusWohlsen	74	0.339
french	0.004	piersmorgan	96	0.348	facebook	0.005	marissamayer	76	0.334
today	0.004	maddow	72	0.314	yahoo	0.005	harrymccracken	43	0.264
news	0.004	BuzzFeedBen	21	0.249	app	0.005	dens	33	0.209
police	0.003	MichaelSteele	81	0.244	google	0.004	nickbilton	89	0.204

Table 2: Topics - document distribution θ

User ID 02		User ID 12		User ID 22		User ID 53		User ID 75		User ID 83		User ID 115		User ID 132	
Topic	Prob.	Topic	Prob.	Topic	Prob.	Topic	Prob.	Topic	Prob.	Topic	Prob.	Topic	Prob.	Topic	Prob.
10	0.090	3	0.504	10	0.506	17	0.733	19	0.526	8	0.249	18	0.151	0	0.144
16	0.072	19	0.039	3	0.036	1	0.017	2	0.029	0	0.084	6	0.060	16	0.076
12	0.065	10	0.036	19	0.034	18	0.016	3	0.029	11	0.06	11	0.060	12	0.070
18	0.064	15	0.035	14	0.031	13	0.016	5	0.028	7	0.045	0	0.058	18	0.057
0	0.058	13	0.032	4	0.03	11	0.015	105	0.028	12	0.043	9	0.054	15	0.050

5. APPLICATION

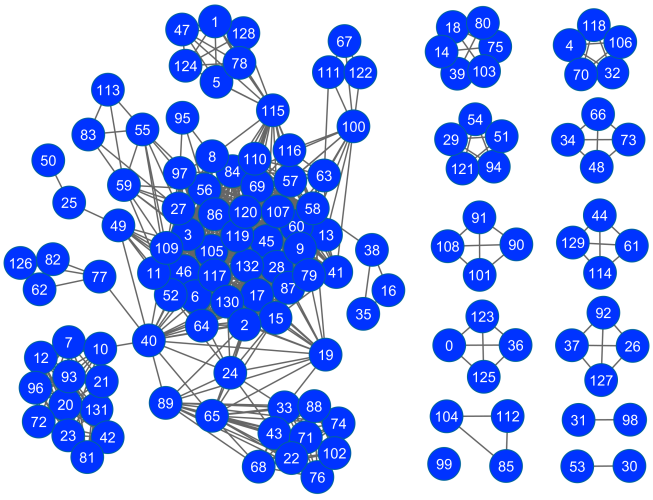

Figure 7: Network for 133 users with two relationships based on Hellinger distance ($distance \leq 0.15$) and Major topic (probability ≥ 0.15).

The microblogging service twitter has become one of the major micro-blogging websites, where people can create and exchange content with a large audience. In this section, we apply the MCPTP algorithm for clustering a set of users around their center of interests. We have targeted 133 users and gathered their tweets in 133 document. We have cleaned them and run the *LDA Gibbs Sampling* algorithm to define the topics distribution of each document and words distribution of each topic. We have used then, the *MCPTP* algorithm to automatically detect the different communities for clustering users. We present in the following, the latter steps in more details.

5.1 Data collection

Twitter is a micro-blogging social-media website that provides a platform for the users to post or exchange text messages of 140 characters. Twitter provides an API that allows easy access to anyone to retrieve at most 1% sample of all the data by providing some parameters. In spite of the 1% restriction we are able to collect large data sets that contain enough text information for topic modeling as it is shown in [9].

The data set contains tweets from the 133 famous and most followed public accounts. We have chosen these accounts because they are characterized by the heterogeneity of the tweets they posts. The followers that they aim to reach comes from different interest area (i.e. politics, technology, sports, art, etc ..). We used the API provided by twitter to collect the messages of 140 characters between January and February, 2015. We gathered all the tweets from a user into a document.

5.2 Data pre-processing

Social media data and mainly twitter data is highly unstructured: typos, bad grammar, presence of unwanted content like: humans expressions (happy, sad, excited, ...), URLs, stop words (the, a, there, ...). To get good insights and to build better algorithms it is essential to play with clean data. The pre-processing step get the textual data clean and ready as an input for the MCPTM algorithm.

5.3 Topic modeling results

After collecting and pre-processing data, we obtained data with 133 documents, 158,578 words in corpus which averages 1,192 words per document and 29,104 different words in the vocabulary, we run LDA Gibbs Sampling from algorithm 1 (Fig. 2) and received the output with two matrices, the document-topic distribution matrix θ and the topic-term distribution matrix ϕ. We presented in table 1 two topics from list of 20 topics that we have computed with our LDA implementation. A topic is presented with a distribution of words. For each topic we have a list of users. Each user is identified with an ID from 0 to 132 and is associated to a topic with an order of probabilities. The two lists of probabilities in topic 3, 10 are extracted respectively from θ and ϕ distributions. The topic 3 and topic 10 are of particular interest due to the important number of users that are related to them. Topic 3 is about the terrorist attack that happened in Paris and topic 10 is about the international Consumer Electronics Show (CES). The both events happened at the same time when we collected our data from

Table 3: Classification documents based on their major topic

Major Topic	$prob \geq 0.3$	$0.15 < prob < 0.3$
Topic 0	112,85,104	-
Topic 1	44,129,114	61
Topic 2	101,108,91	90
Topic 3	42,23,12,7,20,131,96,72	21,81,93,10
Topic 4	125,36,123,0	-
Topic 5	82,126	62
Topic 6	127,37,26	92
Topic 7	118,106,32	70,4
Topic 8	113	83,55,59
Topic 9	67,122	111,100
Topic 10	22,102,88,71,74,68,76	43,89,33,65
Topic 11	54,51,121	29,94
Topic 12	50	12
Topic 13	16,35	38
Topic 14	31,98	-
Topic 15	66,73,34,	48
Topic 16	99	-
Topic 17	53,30	-
Topic 18	47,128,1,124,5	78,115
Topic 19	14,80,39,75,18,103	-
None	**remaining users (with probability < 0.15)**	

Twitter. We note that we have more users for these topics compared to the other. We can conclude that these topics can be considered as hot topics at this moment.

Due to the lack of space, we could not present in details all the topics with their distribution of words and all topic distributions of documents. Therefore, we presented eight topic distributions θ_i (sorted by probability) of eight users in the table 2. A high probability value of θ_{ij} indicates that document i is closely related to topic j. Hence, user ID 12 is closely related to topic 3, user ID 22 closely related to topic 10, etc. In addition, documents with low θ_{ij} values over all the topics are noisy documents that belong to none of the topics. So, there are no major topic in user ID 02 and user ID 132 (the max probability < 0.15). We showed in the table 3 a classification of documents based on their major topics in two levels along their probability, the documents with max probability < 0.15 is considered noisy documents and clustered in the same cluster. The clustering of users was done manually. We present in the next subsection the results of the clustering of the algorithm that we have developed in our article.

5.4 Results from the k-means algorithm using Hellinger distance

After receiving the document-topic distribution matrix θ from LDA Gibbs Sampling, we used k-means algorithm with Hellinger distance to cluster users. The table 4 presents the result from k-means algorithm using Hellinger distance with number of clusters k=13 and random centroids. Based on the mean value of each cluster, we defined the major topic related to the clusters and attached these values in the table. We notice that different choices of initial seed sets can result in very different final partitions.

5.5 Results from the MCPTM algorithm

After getting the results (e.g table 2) from our LDA implementation, we defined two relations between two docements, the first based on their major topic R_{MTP} and the second based their Hellinger distance R_{d_H}. We then built the weak pseudoclosure with these relations and applied to compute pseudoclosure distance and the minimal closed subsets. With this pseudoclosure distance, we can use MCPTP algorithm to cluster set of users with multi-relationships.

Table 4: Result from k-means algorithm using Hellinger distance

Cluster	Users	Major Topic
1	67, 111, 122	TP 9 (0.423)
2	34, 48, 66, 73	TP 15 (0.315)
3	10, 22, 33, 43, 65, 68, 71, 74, 76, 88, 89, 98, 102	TP 10 (0.305)
4	26, 92	TP 6 (0.268)
5	16, 35, 44, 90, 91, 101, 108, 114, 129	TP 2 (0.238)
6	4, 32, 70, 106, 118	TP 7 (0.345)
7	37, 127	TP 6 (0.580)
8	14, 18, 39, 75, 80, 103	TP 19 (0.531)
9	1, 5, 47, 78, 124, 128	TP 18 (0.453)
10	30, 53	TP 17 (0.711)
11	7, 12, 20, 21, 23, 42, 72, 81, 93, 96, 131	TP 3 (0.409)
12	0, 31, 36, 82, 123, 125	TP 4 (0.310)
13	remaining users	None

The figure 8 shows the number of elements of minimal closed subsets with different thresholds p_0 for R_{MTP} and d_0 for R_{d_H}. We used this information to choose the number of clusters. For this example, we chose $p_0 = 0.15$ and $d_0 = 0.15$ i.e user i connect with user j if they have the same major topic (with probability ≥ 0.15) or the Hellinger distance $d_H(\theta_i, \theta_j) \leq 0.15$. From the network (figure 7) for 133 users built from the weak pseudoclosure, we chose the number of clusters $k = 13$ since the network has 13 components (each component represents an element of the minimal closed subset). We used inter-pseudoslosure distance to compute initial centroids and received the result:

$$[0, 52, 4, 14, 26, 29, 30, 31, 34, 44, 85, 90, 99]$$

The table 5 presents the results of MCPTP algorithm and k-means algorithm using Hellinger distance. We notice that there are almost no difference between the results from two methods when using the number of clusters k and initial centroids above. By looking deeply into each cluster, we saw that the accuracy of these results are better than the result when we chose centroids randomly.

Table 5: Result from k-means algorithm using Hellinger distance and MCPTP

	K-means & Hellinger		MCPTP Algorithm	
Cluster	Users	Topic	Users	Topic
1	0,36,123,125	TP 4 (0.457)	0,36,123,125	TP 4
2	4,32,70,10,118	TP 7 (0.345)	4,32,70,10,118	TP 7
3	14,18,39,75,80,103	TP 19 (0.531)	14,18,39,75,80,103	TP 19
4	26,37,92,127	TP 6 (0.424)	26,37,92,127	TP 6
5	29,51,54,94,121	TP 11 (0.345)	29,51,54,94,121	TP 11
6	30,53	TP 17 (0.711)	30,53	TP 17
7	31	TP 14 (0.726)	31,98	TP 14
8	34,48,66,73	TP 15 (0.315)	34,48,66,73	TP 15
9	44,61,114,129	TP 1 (0.413)	44,61,114,129	TP 1
10	85,104,112	TP 0 (0.436)	85,104,112	TP 0
11	67,90,91,101,108	TP 2 (0.407)	90,91,101,108	TP 2
12	99	TP 16 (0.647)	99	TP 16
13	remaining users	None	remaining users	None

We saw that the largest component in users network (fig. 7) has many nodes with weak ties. This component represents the cluster 13 (remaining users) with 89 elements. Hence, we used k-means algorithm with Hellinger distance for clustering this group with number of cluster $k = 9$, centroids:

$$[23, 82, 113, 67, 22, 50, 16, 47, 2]$$

and showed the result in the table 6.

The idea of using pretopology theory for k-means clustering has been proposed by [12]. In this paper, the authors

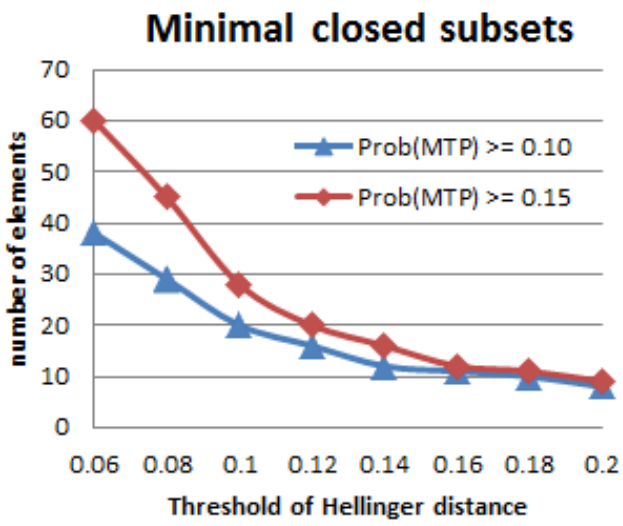

Figure 8: Number of elements of Minimal closed subsets with difference thresholds p_0 for R_{MTP} and d_0 for R_{d_H}.

Table 6: Result from k-means algorithm using Hellinger distance for cluster 13 (89 users)

Cluster	Uesrs	Major Topic
13.1	7, 12, 20, 21, 23, 42, 72, 81, 93, 96, 131	TP 3 (0.409)
13.2	62, 77, 82, 126	TP 5 (0.339)
13.3	27, 55, 59, 83, 113	TP 8 (0.218)
13.4	67, 111, 122	TP 9 (0.422)
13.5	22, 33, 43, 65, 68, 71, 74, 76, 88, 89, 102	TP 10 (0.330)
13.6	50	TP 12 (0.499)
13.7	16, 35	TP 13 (0.576)
13.8	1, 5, 47, 78, 124, 128	TP 18 (0.453)
13.9	remaining users	None

proposed the method to find automatically a number k of clusters and k centroids for k-means clustering by results from minimal closed subsets algorithm and also proposed to use pseudoclosure distance constructed from the relationships family to examine the similarity for both numeric and categorical data. The authors illustrated the method with a toy example about the toxic diffusion between 16 geographical areas using only one relationship. Our work extended this method in two dimensions: firstly, we exploited this idea in document clustering and integrated structural information from LDA; secondly, we showed that pretopology theory can apply for multi-criteria clustering by defining pseudodistance build from multi-relationships. In our paper, we clustered documents by using two criteria: one based on the major topic of document (qualitative criterion) and the other based on Hellinger distace (quantitative criterion). Our application in Twitter also proposed a method to construct a network from multi-relations network by choosing the set of relations and then applying strong or weak pretopology.

6. CONCLUSION

The major implication intended by this article is that the number of clusters and the chosen criterias for grouping the document is closely bounded with the accuracy of the clustering results. The method presented here can be considered as a pipeline where we associate Latent Dirichlet Allocation (LDA) and pseudoclosure function. LDA is used to estimate topic-distribution of each document in corpus and pseudoclosure function to connect documents with multi-relations built from their major topics or Hellinger distance. With this method both quantitative data and categorical data are used, allowing us to have a multi-criteria clustering. We have presented our contribution by applying it on microblogging posts and have obtained good results. In future works, we want to test these results on a more important scale where we will need to parallelize the developed algorithms.

7. REFERENCES

[1] Z. Belmandt. *Basics of Pretopology*. Hermann, 2011.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[3] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[4] G. Heinrich. Parameter estimation for text analysis. Technical report, 2004.

[5] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[6] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[7] V. Levorato and M. Bui. Modeling the complex dynamics of distributed communities of the web with pretopology. *Proceedings of the 7th International Conference on Innovative Internet Community Systems*, 2007.

[8] C. D. Manning and P. Raghavan. An Introduction to Information Retrieval, 2009.

[9] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. *arXiv:1306.5204 [physics]*, June 2013. arXiv: 1306.5204.

[10] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press, 2004.

[11] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[12] N. K. Thanh Van Le and M. Lamure. A clustering method associating pretopological concepts and k-means algorithm. *The 12th International Conference on Applied Stochastic Models and Data Analysis*, 2007.