Non-disjoint grouping of text documents based Word Sequence Kernel

Chiheb-Eddine Ben N'Cir*, Afef Zenned**,Nadia Essoussi***

*LARODEC, ISGT, University of Tunis
chiheb.benncir@isg.rnu.tn
**LARODEC, ISGT, University of Tunis
afef.zenned@gmail.com
***LARODEC, ISGT, University of Tunis
nadia.essoussi@isg.rnu.tn

Abstract. This paper deals with two issues in text clustering which are the detection of non disjoint groups and the representation of textual data. In fact, a text document can discuss several themes and then, it must belong to several groups. The learning algorithm must be able to produce non disjoint clusters and assigns documents to several clusters. The second issue concerns the data representation. Textual data are often represented as a bag of features such as terms, phrases or concepts. This representation of text avoids correlation between terms and doesn't give importance to the order of words in the text. We propose a non supervised learning method able to detect overlapping groups in text document by considering text as a sequence of words and using the Word Sequence Kernel as similarity measure. The experiments show that the proposed method outperforms existing overlapping methods using the bag of word representation in terms of clustering accuracy and detect more relevant groups in textual documents.

1 Introduction

Text clustering is an important application within the information Retrieval field (IR). It aims to group similar documents in the same cluster, while dissimilar documents must belong to different clusters without using any predefined categories. This definition can be a crucial issue in many real life applications of text clustering where a document needs to be assigned to more than one group. This issue arises naturally because a document can discuss several topics and can belong to several themes. For example, a newspaper article concerning the participation of a soccer in the release of an action film can be grouped with both of the categories Sports and Movies.

Many clustering methods have been proposed to solve the problem of the detection of non disjoint groups in data. This kind of application is refereed as overlapping clustering (Diday, 1984), (Fellows et al., 2011). Our works concerns the detection of groups based k-means algorithm. Existing overlapping methods, when applied to text document clustering (Cleuziou, 2008), use the Vector Space Model (VSM) representation for the set of documents. This representation is based on the assumption that relative position of tokens are irrelevant leading to the loss of correlation with adjacent words and to the loss of information regarding word positions. The loss of information and the loss of correlation between adjacent words influence the quality of obtained clusters.

We propose in this paper, to use a structured model for text representation when detecting overlapping groups based on sequences of words. This representation, takes into account information regarding words positions and has the advantage of being more language independent. We propose an overlapping method able to detect relevant groups in textual data based on Word Sequence Kernel as a similarity measure.

This paper is organized as follows: Sect.2 presents the Word Sequence Kernel and existing overlapping clustering methods, then Sect.3 presents the KOKM based WSK method that we propose. Experiments on different data sets are described and discussed in Sect.4. Finally, Sect.5 presents conclusion and future works.

2 Background

In the Vector Space Model (VSM), each text document is represented by a vector of tokens (words) where the size of vector is determined by the number of different tokens in all documents D. Each documents d_j will be transformed into a vector : $d_j = (w_{1j}, w_{2j}, ..., w_{|T|_j})$, where T is the whole set of terms $T = (t_1, ..., t_{|T|})$ (or tokens) which appears at least once in the corpus ($|T|$ is the size of the vocabulary), and w_{kj} represents the weight (frequency or importance) of the term t_k in the document d_j. Documents whose vectors are close to each others based on tokens frequencies are considered to be similar in content. This representation is based on the assumption that relative position of tokens has a little importance leading to the loss of correlation with adjacent words and leading to the loss of the information regarding word positions.

The "n-grams" representation of text solves the problem of the loss of information regarding word positions by considering text document as sequences of n consecutive characters (syllables or words). The whole set of n-grams is obtained by the extraction of all possible ordered subsequences of consecutive n characters (syllables or words) along the text. This representation of text leads to a high dimensional features of subsequences representing the text document. This problem is solved in information retrieval tasks by using kernel machines over sequences of text. Many kernels known as String Kernel such as n-grams Kernel (Leslie et al., 2002), String Subsequence Kernel (Lodhi et al., 2001) and Word Sequence Kernel (Cancedda et al., 2003) are proposed in literature. These kernels return the inner product between documents mapped into a high dimensional feature space. This inner product is computed without explicitly computing feature vectors.

2.1 WSK : Word Sequence Kernel

WSK is defined as an extension of the String Sequence Kernel (SSK) proposed by Lodhi et al. (2001) which measure the similarity between two sentences or two documents based on the number of sequences of characters shared between them. Cancedda et al. (2003) extends the

SSK kernel and proposes the WSK kernel which measure the similarity between two sequences based *word* rather then *character*.

Let Σ the alphabet which consists in the set of words that exist in all documents, let $S = s_1 s_2 s_3 ... s_{|s|}$ the sequence of words with $|S|$ is the length of S, let $u = s[i]$ a subsequence of S with $s[i] = s_{i_1}..s_{i_j}..s_{i_n}$ where s_{i_1} and s_{i_j} in this subsequence are not necessarily contiguous in S, the feature mapping ϕ for the sentences s in the feature space is given by defining ϕ_u for each $u \in \Sigma^n$ as:

$$\phi_u(s) = \sum_{i:u=s[i]} \lambda^{l(i)}, \tag{1}$$

where $l(i)$ is the length of subsequence $s[i]$ in S with $l(i) = i_n - i_1 + 1$ and λ is the decay factor used to penalize non contiguous subsequences. These features measure the number of occurrences of subsequence u in the sentences s weighting them according to their lengths. So, given two strings s_1 and s_2, the inner product of the feature vectors is obtained by computing the sum over all common subsequences :

$$\begin{aligned} K_n(s_1, s_2) &= \sum_{u \in \Sigma^n} \phi_u(s_1)\phi_u(s_2) \\ &= \sum_{u \in \Sigma^n} \sum_{i:u=s_1[i]} \sum_{j:u=s_2[j]} \lambda^{l(i)+l(j)}. \end{aligned} \tag{2}$$

The objective of this representation is to keep information regarding words positions and to keep linguistic meaning of the terms when using ordered words as atomic units. For example, the terms "son-in-law" mean a special meaning that can be lost if it is broken. In addition, the number of features per document is reduced because it uses sequences of words rather than sequences of characters.

2.2 Existing Overlapping methods based k-means algorithm

Existing k-means based methods extend fuzzy and possibilistic clustering methods to produce overlapping clusters. Example of these methods are the fuzzy c-means (Bezdek, 1981) and the possibilistic c-means (Krishnapuram and Keller, 1993) methods. These methods need a post-processing treatment to generate hard and overlapping clusters by thresholding clusters memberships. More recent methods are based on the adaptation of k-means algorithms to look for optimal covers (Cleuziou, 2008), (Cleuziou, 2009), (BenN'Cir et al., 2010). As opposed to fuzzy and possibilistic k-means, these methods produce hard overlapping clusters and does not need any post processing treatment. The criteria optimized by these methods look for optimal overlapping groups.

A recent proposed method referred as Kernel overlapping k-means (KOKMϕ) (BenN'Cir and Essoussi, 2012), extends kernel k-means to detect non disjoint and non-linearly separable clusters. By an implicit mapping of the data from an input space to a higher, possibly infinite, feature space, KOKMϕ looks for separation in feature space and solves the problem of overlapping clusters with non-linear and non-spherical separations.

Given the set of observations $X = \{x_i\}_{i=0}^N$ with $x_i \in R^d$, and N is the number of observations. Let C the number of covers and $\phi(x_i)$ the representation of the observation x_i in a hight dimension space with a non linear transformation $\phi : x_i \mapsto \phi(x_i) \in F$. The KOKM$\phi$

method introduces the overlapping constraint (an observation can belong to more than one cluster) in the objective function which minimizes a local error on each observation defined by the distance between the observation and it's $image$ in the feature space :

$$J(\{\pi_c\}_{c=1}^{C}) \quad = \quad \sum_{x_i \in X} \|\phi(x_i) - im(\phi(x_i))\|^2, \tag{3}$$

where $im(\phi(x_i))$ is the $image$ of the observation x_i and is defined by the gravity center of clusters prototypes to which x_i belongs. If observation x_i is assigned to only one cluster, the $image$ is equivalent to the representative of the following cluster. The objective function is computed without explicitly performing the non linear mapping ϕ using the function of the Kernel $K_{ij} = \phi(x_i).\phi(x_j)$ evaluating the dot product in feature space between x_i and x_j:

$$J(\{\pi_c\}_{c=1}^{C}) \quad = \quad \sum_{x_i \in X} [K_{ii} - \frac{2}{L_i} \sum_{c=1}^{C} P_{ic} \cdot K_{im_c} + (\frac{1}{L_i})^2 \sum_{c=1}^{C} \sum_{l=1}^{C} P_{ic} P_{il} \cdot K_{m_c m_l}]. \tag{4}$$

3 Proposed solution : KOKM based WSK

To detect non disjoint groups from sequential text documents, we propose an overlapping method refereed as "KOKM based WSK" using WSK as a similarity measure between structured documents. Given a set of documents $D = \{d_q\}_{q=1}^{|D|}$ where each document d_q is defined in the feature space by the u coordinate $\phi_u(d_q)$ which measures the number of occurrences of subsequence u in the document d_q weighted according to it's lengths. The proposed method consists on the minimization of a local error on each document, where the local error is defined by the kernel distance between the document and it's image. The objective function of KOKM based WSK is described by:

$$J(\{\pi_c\}_{c=1}^{C}) = \sum_{d_q \in D} \|\phi(d_q) - im(\phi(d_q))\|^2 . \tag{5}$$

where C is the number of overlapping clusters and $im(\phi(d_q))$ is the image of document d_q defined by the gravity center of clusters prototypes to which document d_q belongs:

$$im(\phi(d_q)) = \frac{\sum_{c=1}^{C} P_{qc} \cdot \phi(m_c)}{\sum_{c=1}^{C} P_{qc}}, \tag{6}$$

with P_{qc} is a binary variable indicating the membership of document d_q in cluster c and m_c is the prototype of the cluster c in the feature space. Using the defined WSK Kernel in equation 2 and using the Kernel Trick, the objective function is performed as follows:

$$\begin{aligned} J(\{\pi_c\}_{c=1}^{C}) &= \sum_{d_q \in D} \big[\sum_{u \in \Sigma^n} \phi_u(d_q)\phi_u(d_q) - \frac{2}{L_q}\sum_{c=1}^{C} P_{qc} \cdot \sum_{u \in \Sigma^n} \phi_u(d_q)\phi_u(d_{m_c}) + \\ & \quad (\tfrac{1}{L_q})^2 \sum_{c=1}^{C}\sum_{l=1}^{C} P_{qc}P_{ql} \cdot \sum_{u \in \Sigma^n} \phi_u(d_{m_c})\phi_u(d_{m_l})\big] \\ &= \sum_{d_q \in D} \big[K_n(d_q,d_q) - \frac{2}{L_q}\sum_{c=1}^{C} P_{qc} \cdot K_n(d_q,d_{m_c}) + \\ & \quad \tfrac{1}{L_q}^2 \sum_{c=1}^{C}\sum_{l=1}^{C} P_{qc}P_{ql} \cdot K_n(d_{m_c},d_{m_l})\big], \end{aligned} \tag{7}$$

where $L_q = \sum_{c=1}^{C} P_{qc}$, Σ^n is the set of all possible ordered word-sequences of length n and d_{m_c} is the prototype of cluster c.

The minimization of the objective function is performed locally by iterating two principal steps: the first step concerns the computation of clusters prototypes where each prototype is defined by the document that minimizes all distances with others documents belonging to the same group. This computation of prototypes is performed in the feature space where the document d_{m_c} representing prototype of cluster c is computed as follows:

$$\begin{aligned} d_{m_c} &= \min_{q \in \pi_c} \frac{\sum_{j \in \pi_c, j \neq q} w_j . \sum_{u \in \Sigma^n} ||\phi_u(d_q) - \phi_u(d_j)||^2}{\sum_{j \in \pi_c, j \neq q} w_j} \\ &= \min_{q \in \pi_c} \frac{\sum_{j \in \pi_c, j \neq q} w_j [K_n(d_q,d_q) - 2.K_n(d_q,d_j) + K_n(d_j,d_j)]}{\sum_{j \in \pi_c, j \neq q} w_j}, \end{aligned} \tag{8}$$

where w_j is a weight of the Kernel distance between document d_q and document d_j depending on the number of groups to which document d_j belongs. This weight is more important if document d_j belongs to more than one cluster to take into account that overlapping documents d_j have a small influence in determining cluster prototype as well as the number of assignment increases.

The second step concerns the multi assignment of documents to one or several clusters. This step is performed using an heuristic that explores the combinatorial sets of possible assignments. The heuristic consists, for each document, in sorting clusters from closest to the farthest, then assigning the document in the order defined while assignment minimizes the local error defined in the objective function and then it minimizes the hole objective function. The stopping rule of KOKM based WSK algorithm is characterized by two criteria: the maximum number of iterations or the minimum improvement of the objective function between two iterations. The main algorithm of KOKM based WSK is described as follows:

Algorithm 1 KOKM based WSK $(D, C, t_{max}, \varepsilon) \rightarrow \{\pi_c\}_{c=1}^{C}$

```
Require: D: set of Documents,
    $t_{max}$: maximum number of iterations,
    $\varepsilon$: the minimum improvement in the objective function,
    C: number of groups.
Ensure:
  Initialize prototypes of clusters with a Random documents, Assign documents and derive
  value of the objective function $J_0(\{\pi_c\}_{c=1}^{C})$ in iteration 0 using equation 7.
  Compute prototypes of clusters using equation 8.
  Assign documents to one or several clusters.
  Compute $J_t(\{\pi_c\}_{c=1}^{C})$ using equation 7.
  if ($t < t_{max}$ and $J_{t-1}(\{\pi_c\}_{c=1}^{C})$ - $J_t(\{\pi_c\}_{c=1}^{C}) > \varepsilon$) then
      go to step 2.
  else
      Return the clusters memberships $\{\pi_c\}_{c=1}^{C}$ in iteration $t$.
  end if
```

4 Experiments and Discussions

Experiments were performed on computer with 4 GB RAM and 2.1 GHZ Intel Core 2 duo processor. Data are preprocessed by removing a stop words. The Vector Space Model representation of each data set is built using the "WEKA text preprocessing module" where the frequencies of occurrence of words is computed using the $TF * IDF$ technique. For the implementation of Word Sequence Kernel, we use the recursive definition of WSK defined by Cancedda et al. (2003) based on the dynamic programming technique. The advantage of this implementation is to reduce the time complexity and to perform WSK without implicitly extracting word sequences. The time complexity of computing WSK kernel between two documents d_1 and d_2 is reduced to $O(n|d_1||d_2|)$ where n is the length of the used subsequence and $|d_i|$ is the number of words in document d_i. The computational complexity of the KOKM based WSK method is evaluated to $O(N.C^2.N_c)$ where N is the number of documents, C is the number of clusters and N_c is the maximal number of documents in each clusters.

Experiments were conducted on two overlapping textual data sets which are respectively Reuters [1] and Ohsumed [2] datasets. We used a subset of Reuters composed of 76 documents and a subset of Ohsumed composed of 83 documents. Each document in each data set is labeled by one or many labels from a set of 5 categories where each category contains 20 documents. Results are compared according to three external validation measures: *Precision*, *Recall* and *F-measure*. These validation measures attempt to estimate whether the prediction of categories is correct with respect to the underlying true categories in the data. For each data set, the number of groups is set to the number of underlying categories in the data set.

Table 1 and Table 2 report average scores and the standard deviation variation of Precision, Recall and F-measure on ten runs using OKM and KOKMϕ methods based on VSM representation compared to the proposed method KOKM based WSK (we fix $n = 2$ the length of

[1] cf. http://kdd.ics.uci.edu/databases/reuters-transcribed/reuters-transcribed.html

[2] cf. http://disi.unitn.it/moschitti/corpora/ohsumed-first-20000-docs.tar.gz

	With stemming			Without stemming		
Methods	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
OKM	0,275±0,01	**0,968±0,03**	0,429±0,01	0,275±0,01	**0,968±0,03**	0,429±0,01
KOKMϕ (Linear)	0,275±0,01	0,955±0,04	0,427±0,02	0,275±0,01	0,958±0,04	0,428±0,01
KOKMϕ (Polynomial)	0,275±0,01	0,955±0,04	0,427±0,01	0,275±0,01	0,958±0,04	0,428±0,01
KOKMϕ (RBF σ=10)	0,274±0,01	0,965±0,03	0,426±0,02	0,274±0,01	**0,968±0,03**	0,427±0,02
KOKMϕ (RBF σ=10^8)	0,275±0,01	0,955±0,05	0,427±0,02	0,275±0,01	0,958±0,04	0,428±0,01
KOKM based WSK	**0,499±0,04**	0,670±0,12	**0,569±0,04**	**0,458±0,05**	0,698±0,02	**0,553±0,04**

TAB. 1 – *Comparison between OKM and KOKMϕ based VSM representation with KOKM based WSK on Reuters Dataset.*

	With stemming			Without stemming		
Methods	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
OKM	0,274±0,03	0,799±0,36	0,396±0,04	0,262±0,04	0,761±0,32	0,378±0,02
KOKMϕ (Linear)	0,297±0,10	0,798±0,36	0,417 ±0,11	0,297±0,10	0,795±0,36	0,416±0,11
KOKMϕ (Polynomial)	0,297±0,10	0,798±0,35	0,417±0,10	0,297±0,10	0,795±0,36	0,417±0,11
KOKMϕ (RBF σ=10)	0,248±0,01	**0,980±0,02**	0,396±0,01	0,248±0,01	**0,983±0,02**	0,396±0,01
KOKMϕ (RBF σ=10^8)	0,262±0,04	0,835±0,40	0,385±0,03	0,260±0,04	0,840±0,40	0,383±0,03
KOKM based WSK	**0,308±0,03**	0,696±0,08	**0,421±0,02**	**0,312±0,02**	0,641±0,06	**0,420±0,03**

TAB. 2 – *Comparison between OKM and KOKMϕ based VSM representation with KOKM based WSK on Ohsumed Dataset .*

word sequences and λ=0.9 the value of the decay factor). Results are compared with and without stemming. For each run, all methods are computed with the same initialization of seeds to guarantee that all methods have the same experimental conditions. Values in bold correspond to the best obtained scores.

The F-measure obtained with KOKM based WSK is characterized by a high value compared to overlapping methods based VSM representation. The improvement of F-measure is induced by the improvement of precision. For example, in Reuters data set the obtained precision using KOKM based WSK is 0.458 while using KOKMϕ and OKM methods the max obtained precision is 0.275. The improvement of precision is realized with and without stemming. Recalls obtained with KOKMϕ and OKM methods are characterized by a high values (the Recall obtained with OKM in Reuters data set is equivalent to 0.968). These high values of recall is explained by the way that OKM and KOKMϕ assign observations to all clusters because of the diagonal dominance problem. For example, in Reuters data set, where the dimensionality of the VSM matrix is very sparse (1482 words), OKM and KOKMϕ have the issue of diagonal dominance and therefore assigns each observation to all clusters.

Obtained results prove the theoretical finding that considering text as a sequence of words improves clustering accuracy compared to VSM representation. The meaning of natural languages depends on the word sequences, and the frequent word sequences can provide compact and valuable information about documents structures. In fact, methods based kernel function (Bag of Word Kernel or Word Sequence Kernel) outperform non kernel method. These results prove that looking for separation between clusters in a feature space is better than looking for separation in input space. Separability between documents can be improved when documents are mapped to a feature space.

5 Conclusion

We proposed in this paper the KOKM based WSK method which is able to detect non-disjoint groups from textual sequential documents based on Word Sequence kernel as a similarity measure. Detecting overlapping groups by considering text as a sequence of words improves quality of obtained groups compared to the VSM representation of text. Preliminaries obtained results on Reuters and Ohsumed data sets prove the efficiency of the proposed method compared to overlapping methods using the VSM representation.

This proposed method can be applied for many other application domains where textual data needs to be assigned to more than one cluster. We plan to conduct experiments in other real overlapping data sets where sequences of text are more relevant than sequences in Reuters and Ohsumed datasetd such as in the detection of groups in textual biological data.

References

BenN'Cir, C. and N. Essoussi (2012). Overlapping patterns recognition with linear and non-linear separations using positive definite kernels. *International Journal of Computer Applications (IJCA).*

BenN'Cir, C., N. Essoussi, and P. Bertrand (2010). Kernel overlapping k-means for clustering in feature space. In *International Conference on Knowledge discovery and Information Retrieval KDIR*, Valencia, SPA, pp. 250–256. SciTePress Digital Library.

Bezdek, J. C. (1981). Pattern recognition with fuzzy objective function algoritms. *Plenum Press 4(2)*, 67–76.

Cancedda, N., E. Gaussier, C. Goutte, and J. Renders (2003). Word-sequence kernels. *Journal of Machine Learning Research 3*, 1059–1082.

Cleuziou, G. (2008). An extended version of the k-means method for overlapping clustering. In *International Conference on Pattern Recognition ICPR*, Florida, USA, pp. 1–4. IEEE.

Cleuziou, G. (2009). Okmed et wokm : deux variantes de okm pour la classification recouvrante. *Revue des Nouvelles Technologies de l'Information, CÃl'paduÃĺs-Edition 1*, 31–42.

Diday, E. (1984). Orders and overlapping clusters by pyramids. Technical Report 730, INRIA, France.

Fellows, M. R., J. Guo, C. Komusiewicz, R. Niedermeier, and J. Uhlmann (2011). Graph-based data clustering with overlaps. *Discrete Optimization 8*(1), 2–17.

Krishnapuram, R. and J. M. Keller (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems 1*, 98–110.

Leslie, C. S., E. Eskin, and W. S. Noble (2002). The spectrum kernel: A string kernel for svm protein classification. In *Pacific Symposium on Biocomputing*, pp. 566–575.

Lodhi, H., N. Cristianini, J. Shawe-Taylor, and C. Watkins (2001). Text classication using string kernel. *The Journal of Machine Learning Research 2*, 419–444.

Résumé

Ce travail traite deux problématiques relatives à la classification des données textuelles. La Première problématique concerne la détection des groupes non-disjoints. En effet, un document textuel peut aborder plusieurs thématiques différentes et par la suite il doit appartenir à plusieurs groupes. L'algorithme d'apprentissage doit tenir compte de cette contrainte et doit assigner chaque document à un ou à plusieurs groupes differents. La deuxième problématique concerne la modélisation des données textuelles. Les données textuelles sont souvent modélisées sous formes vectorielles. Cette forme de représentation néglige la corrélation entre les termes et ne donne aucune importance à l'ordre d'apparitions des mots dans le texte. Nous proposons une méthode de classification non supervisée capable de détecter des groupes avec recouvrements en modélisant le texte sous forme de séquences de termes. Le noyau WSK (Word Sequence Kernel) est utilisé comme mesure de similarité entre les sequences de termes. Les expérimentations réalisées montrent la performance de la méthode proposée par rapport aux méthodes recouvrantes existantes qui utilisent la modélisation vectorielle.