Document Representation Based on Maximal Frequent Sequence Sets

Edith Hernández-Reyes, J. Fco. Martínez-Trinidad, J.A. Carrasco-Ochoa,
and René A. García-Hernández

National Institute for Astrophysics, Optics and Electronics
Luis Enrique Erro No.1 Sta. Ma. Tonantzintla, Puebla, México C. P. 72840
{ereyes, fmartine, ariel, renearnulfo}@inaoep.mx

Abstract. In document clustering, documents are commonly represented through the vector space model as a word vector where the features correspond to the words of the documents. However, there are a lot of words in a document set; therefore the vector size could be enormous. Also, the vector space model does not take into account the word order that could be useful to group similar documents. In order to reduce these disadvantages, we propose a new document representation in which each document is represented as a set of its maximal frequent sequences. The proposed document representation is applied for document clustering and the quality of the clustering is evaluated through internal and external measures, the results are compared with those obtained with the vector space model.

1 Introduction

Currently there is a lot of digital information in the World Wide Web and the amount of digital text documents is growing every day. In consequence, we need automatic methods for organizing these text documents, because if we have organized the text documents it is easier to look for the information that we want.

Document clustering consists in dividing, automatically, a set of documents into different groups [1]. Therefore, the document clustering methods can be seen as tools which allow us to organize documents. Document clustering has been used for tasks such as information retrieval and generating hierarchies of documents [2].

In order to cluster documents, we need to represent them in a suitable way to be compared. Commonly, text documents are represented with the vector space model proposed by Salton in 1975 [3]. This model represents the documents as a word vector in which the features correspond to the words of the documents. However, the vector space model presents some disadvantages such as very high dimensionality, because a document collection can have a huge amount of different words. Another drawback, of the vector space model, is that it does not preserve the word sequential order.

The words are not the only features that can be used in order to represent documents; there is another kind of information that can be extracted from the documents, for example word sequences which appear frequently in the document. But the amount of frequent word sequences could be large; however, the amount of frequent word sequence can be reduced if only the maximal frequent sequences are used.

In this work, we propose a new document representation using the maximal frequent sequences, with this proposed representation each documents is represented as a set of its maximal frequent sequences which reduce the number of features used and preserve part of the word sequential order of the document. In order to test the proposed document representation, some document clustering experiments were done with two document collections: the English document collection Reuters-21578 [4] and the Spanish document collection Disasters. The quality obtained in the document clustering experiments, with the proposed representation, was compared against the one obtained with the vector space model, through internal and external clustering quality measures.

This paper is organized as follows. Section 2 describes the maximal frequent sequences. Section 3 introduces the new document representation. Section 4 gives the methodology used in this works and the experimental results. Finally, in section 5 we present our conclusions and some directions for future work.

2 Maximal Frequent Sequences

The text of a document is expressed by words in a sequential order. Therefore, it could be useful determining the consecutive word sequences that appear frequently in a document. Also, it is possible to determine which of the frequent sequences are not contained in any other frequent sequence *i.e.* which of them are maximal. In this work we will focus in the MFS's because they are a compact representation of the frequent sequences.

Ahonen [5] developed the first algorithm to find sequential patterns in a document collection. Recently, the MFS's have been used by Doucet [6] in the document retrieval task, his algorithm finds the MFS's from a document collection too. In [7] an algorithm to find efficiently the maximal consecutive frequent sequences of words but unlike Ahonen and Doucet algorithms from a single document was proposed.

The maximal frequent sequences are formally defined as follows [7]:

Definition 1. A sequence $P=p_1p_2...p_n$ is a subsequence of a sequence $S=s_1s_2...s_m$, denoted $P \subseteq S$, if there exists an integer $1 \leq i$ such that $p_1=s_i, p_2=s_{i+1}, p_3=s_{i+2}, ..., p_n=s_{i+(n-1)}$.

Definition 2. Let $X \subseteq S$ and $Y \subseteq S$ then X and Y are exclusive if X and Y do not share items *i.e.*, if $(x_n=s_i \text{ and } y_l=s_j)$ or $(y_n=s_i \text{ and } x_l=s_j)$ then $i < j$.

Definition 3. Let T be a text expressed as a sequence, a sequence S is *frequent* in T if it is contained at least β times in T in an exclusive way, where β is the user-specified threshold.

Definition 4. A frequent sequence is *maximal* if it is not a subsequence of any other frequent sequence.

Table 1 presents, as example, two documents and their MFS's with $\beta=2$

The MFS's present some important characteristics. First, they keep the sequential order of the words; it means the MFS's do not lose the sequential order of the text. Second, the length of the MFS's is not previously determined; it is determined by the document content. And third, the MFS's can be obtained independently of the language of the documents.

In this work, the algorithm proposed in [7] was used to obtain the MFS's of each document.

Table 1. MFS's for two documents

d_1 = bank said had provided money market further billion assistance bank afternoon session brings billion bank total help compares revised shortage forecast money market.
MFS's = bank, money market, billion
d_2 = bank billion provided money market late assistance system brings bank total help compares money market latest forecast shortage system today.
MFS's = bank, money market, system

3 Representation Based on MFS Sets

The document representations such as vector of words have some disadvantages like high dimensionality and loss of important information from the sequential order of the original text. In order to reduce these drawbacks we propose a new document representation using MFS's. This representation consists on obtaining the MFS's from each document and use them to represent the document as a set of MFS's. For example, in Table 1 two documents, D_1 and D_2 , are given and they are represented with the set of their MFS's. The first document is represented with the MFS's "bank", "billion" and "money market" and the document D_2 is represented with "bank", "money market" and "system".

In the example from table 1, with the proposed document representation, each document is represented with 3 MFS's while using the vector space model, each document would be represented with a vector of 22 words. Therefore, the number of features which are used in the proposed document representation is less than the number of features used to represent each document with the vector space model. In this sense, the problem of high dimensionality presented in the vector space model could be reduced with the proposed document representation.

It is important highlight that with the MFS's, the sequential order of the words is preserved which could help to distinguish among documents with almost the same words but in different order.

The similarity between two documents represented as MFS sets is based on the intersection of their MFS's and it is obtained with the next expression.

$$sim(d_1, d_2) = \frac{2|MFS(d_1) \cap MFS(d_2)|}{|MFS(d_1)| + |MFS(d_2)|} \tag{1}$$

Where: $MFS(d_j)$ is the set of MFS's of the document j
 $j=1, 2$

Following the idea of the vector space model, the TF-IDF weighting can be used in order to assign a weight to each MFS in the set. Thus the weight w of each MFS S_i for

a document d_j can be computed as the product of its frequency f in d_j and the log of its inverse frequency in the collection C .

$$w_{dj}(S_i) = f_{S_i}(d_j) \log \left(\frac{1}{f_{S_i}(C)} \right) \quad (2)$$

The expression (3) is another way to evaluate the similarity between two documents represented as MFS sets and taking into account the TF-IDF weighting.

$$sim(d_1, d_2) = \frac{2 \left[\sum_{i=S_i \in \{MFS(d_1) \cap SFM(d_2)\}} \left[w_{d_1}(S_i) * w_{d_2}(S_i) \right] \right]}{\left| MFS(d_1) \right| \left| MFS(d_2) \right| + \sum_{i=1}^{\sum w_{d_1}(S_i)} w_{d_1}(S_i) + \sum_{i=1}^{\sum w_{d_2}(S_i)} w_{d_2}(S_i)} \quad (3)$$

Where: $w_{di}(S_i)$ = weight in document d_j of the MFS S_i , $j=1,2$

If the Boolean weight is used, instead of the TF-IDF weighting, in expression (3), this expression will be the same one that the expression (1).

4 Experimentation

In order to test the proposed representation we used the Reuters-21578 and Natural Disasters collection which are written in English and Spanish, respectively. Table 2 and 3 present a description of the data used for the experiments done with Reuters-21578 and Natural Disasters collection. For each experiment, the name of the used

Table 2. Data used for the experiments with the Reuters-21578 collection

	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Classes	Acq earn	Money, acq, earn	Acq, earn, crude	Gold,acq,trade, reserve,earn
Documents	100	120	120	253
Required clusters	2	3	3	5
Total words	1546	2354	2541	4294
Average of MFS's per document	5	6	7	7

Table 3. Data used for the experiments with the Natural Disasters collection

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
Classes	Forest, hurricane	Forest, inundation	Drought, inundation	Forest, earth- quake, inundation	Forest, drought, inundation	Forest, drought, hurricane
Documents	80	80	80	120	120	120
Required clusters	2	2	2	3	3	3
Total Words	4611	4583	4825	5963	6059	4593
Average of MFS's per document	8	11	12	11	22	10

classes, number of documents and the number of required clusters are shown. Also, the number of words and MFS's, from each experiment, are provided. A $\beta=2$ was used in order to extract the MFS's.

4.1 Methodology

In all the experiments, the methodology showed in figure 1 was used. We pre-processed the documents removing punctuation, numbers, special characters and stop words. Then the MFS's from each document were obtained in order to represent each document as a set of its MFS's. For extracting the MFS's, we have used the algorithm described in [7] taking the threshold β equal to 2 since it is the lowest threshold which produced longer MFS's. On the other hand, in order to represent the documents with the vector space model, the bag of words was obtained. Then, with this word set, the documents were represented as word vectors using the Boolean and TF-IDF weighting. After documents were represented with the vector space model and as MFS sets, a clustering algorithm was applied in both document representations.

For documents represented with the vector space model the cosine similarity was used. It is calculated with the next expression:

$$\cos(d_1, d_2) = \frac{(d_1 \cdot d_2)}{\|d_1\| \|d_2\|} \tag{4}$$

For our proposal document representation, the expressions 1 and 3 were used to evaluate the similarity among documents. The used clustering algorithm was the k-means algorithm.

Finally, the clustering was evaluated with internal and external quality measures [8]. Internal measures evaluate the internal cohesion and external separation of the resulting groups without using previous knowledge about the original classes of the collection. In this work, the global similarity and the global silhouette were used as internal measures. For both internal measures, higher values represent better quality of the clusters.

Fig. 1. Methodology of the experiments

It is appropriate to underline that in real clustering problems the original classes are unknown and for this reason the internal measures have been widely used for evaluating clustering. In spite of this, the external measures [8] are employed to evaluate the

clustering quality by comparing the obtained groups against the previously defined classes, which have been determined by a human criterion. In this paper, we evaluated the clustering quality with the external measures total entropy and general F-measure. For general F-measure higher values represent better quality of the clusters and for the total entropy, smaller values represent better quality.

The internal and external measures used in this work are described in table 4.

Table 4. Internal and external measures for evaluating clustering quality

INTERNAL MEASURES	
Silhouette value of the i th document	$s(i) = \frac{\min(AVG_D_BETWEEN(i, k) - AVG_D_WITHIN(i))}{\max(AVG_D_WITHIN(i), AVG_D_BETWEEN(i, k))}$
$AVG_D_BETWEEN(i, k)$: average distance from the i -th document to all documents in other clusters. $AVG_D_WITHIN(i)$: average distance from the i -th document to the others documents in its own cluster.	
Cluster Silhouette	$S_j = \sum_{i=1}^{ C_j } s(i) \quad C_j = \text{number of documents in cluster } C_j$
Global Silhouette	$GlobalSilhouette = \frac{1}{K} \sum_{j=1}^K S_j \quad K = \text{number of clusters}$
$GlobalSimilarity = \frac{\sum_{i=1}^K similarity(C_i)}{K}$	
where: K = number of clusters C_i = cluster i $Similarity(C_i) = \frac{1}{ C_i ^2} \sum_{d \in C_i} \cos(d, d')$	
EXTERNAL MEASURES	
$GeneralFmeasure = \sum_{i=1}^K \left[\frac{ class_i }{N} \max_{j=1..K} \{Fmeasure(i, j)\} \right]$	
where: K = number of classes = number of clusters $ class_i $ = number of documents in the class i N = total number of documents $Fmeasure(i, j) = \frac{2P_{ij}R_{ij}}{P_{ij} + R_{ij}}$ P_{ij} = precision of class i with cluster j R_{ij} = recall of class i with cluster j	
$TotalEntropy = \sum_{j=1}^K \frac{n_j * Entropy_j}{N}$	
where: K = number of clusters n_j = number of documents in cluster i N = total number of documents $Entropy_j = - \sum p_{ij} \log_2(p_{ij})$ p_{ij} = probability that a documents from the cluster j belongs to class i .	

4.2 Results

The results of the experiments for the Spanish and English collections are shown in tables 5 and 6, respectively. In these tables, the first column specifies the used document representation while the second column shows the term weighting used for each representation, the next columns provide the results of each experiment. For each experiment the best results are highlighted.

For the experiments with the Natural Disasters collection, table 5 shows the quality clustering, obtained by the two document representations, evaluated with the internal measures. We can observe that the document representation using MFS's obtained clusters with higher internal cohesion and external separation than the groups obtained with the word vector. This shows that using MFS's as a document representation is a good option. Also, we can observe that the best clustering quality was obtained by document representation using MFS's with Boolean weights, and the smallest quality was obtained by the word vector with TF-IDF term weighting.

Table 5. Clustering quality for the Spanish collection evaluated with the internal measures

GLOBAL SILHOUETTE							
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
Words	Boolean	0,0460	0,0421	0,0651	0,04218	0,05084	0,1623
Words	TF-IDF	0,0186	0,0213	0,0351	0,02487	0,04895	0,0612
MFS's	Boolean	<u>0,4417</u>	<u>0,3144</u>	<u>0,7192</u>	<u>0,2889</u>	<u>0,3976</u>	<u>0,5230</u>
MFS's	TF-IDF	0,1402	0,0848	0,2425	0,1032	0,1354	0,3240
GLOBAL SIMILARITY							
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
Words	Boolean	0,0334	0,0331	0,0167	0,0159	0,0158	0,0618
Words	TF-IDF	0,0156	0,0164	0,0086	0,0077	0,0268	0,0169
MFS's	Boolean	<u>0,1959</u>	<u>0,1863</u>	<u>0,4372</u>	<u>0,1892</u>	<u>0,0853</u>	<u>0,0999</u>
MFS's	TF-IDF	0,0668	0,0596	0,1725	0,0660	0,0268	0,0599

Table 6. Clustering quality for the Spanish collection evaluated with the internal measures

GLOBAL SILHOUETTE					
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Words	Boolean	0,1211	0,1039	0,10314	0,1315
Words	TF-IDF	0,0722	0,0607	0,07701	0,0948
MFS's	Boolean	<u>0,4295</u>	<u>0,3654</u>	<u>0,34480</u>	<u>0,1564</u>
MFS's	TF-IDF	0,1825	0,1595	0,13696	0,0827
GLOBAL SIMILARITY					
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Words	Boolean	0,0484	0,0456	0,0194	0,0251
Words	TF-IDF	0,0228	0,0228	0,0108	0,0136
MFS's	Boolean	<u>0,1619</u>	<u>0,0689</u>	<u>0,1512</u>	<u>0,2173</u>
MFS's	TF-IDF	0,0543	0,0213	0,0482	0,0335

Table 6 presents the results of the clustering quality through internal measures, obtained with the documents in English. The results show that using MFS's for representing documents, the formed groups have high internal cohesion and external separation. The biggest clustering quality was obtained by the document representation using MFS's with Boolean weights, and the smallest quality was obtained by the word vector with TF-IDF term weighting. In both tables, 5 and 6, we can see a great difference of the quality clustering obtained by the MFS's and the word vector is very huge.

In all the experiments, the number of features which were used in the proposed document representation was less than the number of features used to represent each document with the vector space model. For example, in table 2 for the English experiment 3, using the vector space model the documents were represented with 4294 features, words; and using the proposed document representation the average number of MFS's used to represent each document was 7.

In real clustering problems the original classes are unknown and for this reason the internal measures are widely used for evaluating clustering. However, in this paper, the quality clustering was evaluated with external measures too. These measures could be used because we knew the categories of the documents collections that were used.

Table 7 presents the result of the experiments with the Spanish document collection, evaluating the clustering quality with the external measures. Although the document representation using MFS's was not the best in all the experiments it was the best in most of the cases. As we mentioned before, these external measures are based on evaluating the clusters according to a previously defined classification. However we do not always have this information in real problems of document clustering.

Table 7. Quality clustering for the English collection evaluated with the external measures

GENERAL F-MEASURE							
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
Words	Boolean	0,9875	1	0,9624	0,9492	0,9412	0,9499
Words	TF-IDF	1	1	0,9498	0,9492	0,9497	0,9499
MFS's	Boolean	0,9875	1	0,6521	0,9749	0,9751	0,9834
MFS's	TF-IDF	0,9875	1	0,9750	0,9624	0,9916	0,9833
TOTAL ENTROPY							
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6
Words	Boolean	0	0	0,1450	0,0565	0,2141	0,1921
Words	TF-IDF	0	0	0,1962	0,0565	0,2141	0,1526
MFS's	Boolean	0	0	0,9344	0,1450	0,1537	0,1133
MFS's	TF-IDF	0	0	0,1686	0,1962	0,0565	0,1127

Table 8 presents the results of the experiment with the English document collection, evaluating the clustering quality with the external measures. The word vector and document representation using MFS's, both obtained very similar results, and in some cases they were tied.

Table 8. Quality clustering for the English collection evaluated with the external measures

GENERAL F-MEASURE					
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Words	Boolean	<u>0.8886</u>	<u>1</u>	0,90756	0,9123
Words	TF-IDF	<u>0.8886</u>	<u>1</u>	0,90756	0,9123
MFS's	Boolean	0,8782	0,9007	<u>1</u>	<u>1</u>
MFS's	TF-IDF	0,8782	0,8946	<u>1</u>	0,8547
TOTAL ENTROPY					
DOCUMENT REPRESENTATION	WEIGHT	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Words	Boolean	<u>0.4152</u>	<u>0</u>	0,39332	0,4229
Words	TF-IDF	<u>0.4152</u>	<u>0</u>	0,39332	0,4229
MFS's	Boolean	0,4394	0,4099	<u>0</u>	<u>0</u>
MFS's	TF-IDF	0,4394	0,4381	<u>0</u>	0,4678

5 Conclusion

In this paper, we have introduced a new document representation based on MFS's where each document is represented as a set of MFS's. The number of features which are used in the proposed document representation is less than the number of features used to represent each document with the vector space model. In this sense, the problem of high dimensionality presented in the vector space model could be reduced with the proposed document representation. The experiments established that using the maximal frequent as document representation is a good option for document clustering.

Some experiments were made using two documents collections, the results showed that the documents represented as sets of MFS's always obtained clusters with best internal cohesion and external separation. Also, the quality clustering was evaluated with external measures and the document representation using MFS's obtained better quality in most of the experiments.

The objective of the work was to analyze the MFS's performance as document representation for document clustering. However the MFS's have some useful characteristics that could improve even more the document clustering, therefore as future work we propose defining a new way to evaluate the similarity among documents represented by MFS's.

References

1. Zhong Su, Li Zhang, Yue Pan. Document Clustering Based on Vector Quatization and Growing-Cell Structure. Lecture Notes in Computer Science vol. 2718, pp. 326-336, Springer-Verlag 2003.

2. Yoelle S., Fagin, Ronald, Ben-Shaul, Israel Z. y Pelleg, Dan. Ephemeral Document Clustering for Web Applications. IBM Research. Report RJ 10186, 2000.

3. G. Salton, A. Wang, C.S. Yang. A Vector Space Model for Information Retrieval. Journal of the American Society for information Science, pp. 613-620, 1975.

4. [http://www.ics.uci.edu/~kdd/databases/reuters21578/reuters21578.html]

5. Helena Ahonen-Myka. Finding All Maximal Frequent Sequences in Text. Proc. of the ICML99 Workshop on Machine Learning in Text Data Analysis, pp. 11-17, 1999.

6. Antoine Daucet. Advanced Document Description, a Sequential Approach. Thesis PhD. University of Helsinki Finland. 2005.
7. René A. García-Hernández, José Fco. Martínez-Trinidad and Jesús Ariel Carrasco-Ochoa, A Fast Algorithm to Find All the Maximal Frequent Sequences in a Text, 9th Iberoamerican Congress on Pattern Recognition (CIARP'2004), Lecture Notes in Computer Science vol. 3287, pp. 478-486, Springer-Verlag 2004.
8. M. Steinbach, G. Karypis, V. Kumar. A Comparison of Document Clustering Techniques. Proc. Text mining workshop, KDD, 2000.