A New Document Clustering Algorithm for Topic Discovering and Labeling

Henry Anaya-Sánchez¹, Aurora Pons-Porrata¹, and Rafael Berlanga-Llavori²

¹ Center for Pattern Recognition and Data Mining
Universidad de Oriente, Santiago de Cuba, Cuba
{henry,aurora}@cerpamid.co.cu

² Department of Languages and Computer Systems
Universitat Jaume I, Castelló, Spain
berlanga@uji.es

Abstract. In this paper, we introduce a new clustering algorithm for obtaining labeled document clusters that accurately identify the topics of a text collection. In order to determine the topics, our approach relies on both probable term pairs generated from the collection and the estimation of the topic homogeneity associated to term pair clusters. Experimental results obtained over two benchmark text collections demonstrate the utility of this new approach.

Keywords: document clustering, topic discovery, topic descriptions.

1 Introduction

The ever-increasing availability of textual documents has led to a growing challenge for Information Systems to effectively manage and retrieve the information comprised in large collections of texts. Therefore, there exists a great interest in new tools for analyzing and summarizing these collections according to the user's information needs.

Clustering is an unsupervised learning technique commonly used in the process of topic discovery from documents. It is aimed at generating document groups or clusters, each one representing a different topic. However, it is not enough. Users also need to easily find out what a cluster is about for determining at a glance those of their interest.

Unfortunately, common clustering techniques do not provide proper descriptions of the obtained clusters, making them difficult to interpret. The problem of determining the meaning of the clusters is simply left to the user. In this context, our research focuses on two issues: (1) how to discover the topics comprised in a text collection, and (2) how to simultaneously provide a meaningful label for each topic.

The issue of describing automatic detected topics was initially presented in the earlier topic discovering systems such as Scatter/Gather [1] and STC [2]. The former one proposed to describe detected topics with the most frequent words in their corresponding clusters, whereas STC uses frequent word sequences to

describe both documents and groups. Due to the limitations of their underlying clustering algorithms, these approaches are only able to handle either small document collections or short documents (e.g. snippets).

Next generation of Topic Detection Systems (TDS) were indeed devised for large and dynamic document collections such as newswire streams. However, as these approaches mainly rely on existing document clustering techniques, they disregard the description of the detected topics. Recently, in [3] a new TDS is proposed for detecting and describing topics by means of a cluster hierarchy. The aim of the hierarchy is to capture properly the different topic granularities. For describing topics, the Typical Testor theory is applied in order to select the most discriminating frequent words of each cluster. However, this approach has several limitations. Firstly, it requires to build a hierarchy of clusters to detect the proper granularity of the topics. Secondly, descriptions are calculated once the hierarchy is built by selecting a subset of features that discriminates each cluster from the others. To calculate those discriminating subsets use to be computationally expensive at the risk of a combinatorial explosion.

More recently, a series of works aims at obtaining simultaneously both the coverage of a topic and its description by means of a new clustering criterion based on the concept of frequent term sets [4,5,6]. Thus, each frequent term set determines both a cluster and its description. These approaches suffer from four main limitations. First, the importance of a frequent term set only depends on the number of documents that contain it, ignoring thus the weight of each term in the documents. Secondly, each frequent term set is mandatory for the cluster it defines, that is, each cluster document must contain all the terms in the frequent term set. This can be too restrictive since there may be documents about a topic in which these terms do not occur, but term variations or related terms occur instead. Thirdly, terms in a frequent set having low discriminating power can lead to merge actual topics. Finally, these methods need to set up a minimum support, which implies to disregard all the topics with a size below this support.

In this paper, we introduce a clustering algorithm for obtaining labeled document clusters that accurately identify the contents of a text collection. This approach attempts to address the limitations mentioned above. Unlike the previous works, the basis of our approach relies on the probability of term pairs in the document collection and the estimation of the topic homogeneity associated to term pair clusters. These two elements allow us to filter out bad descriptions. Even restricting ourselves to term pairs, experimental results demonstrate the usefulness of this approach.

The rest of the paper is organized as follows. Section 2 presents some basic concepts concerning our proposal. Section 3 describes the clustering algorithm for topic discovering and labeling. In Section 4, some experimental results are shown. Finally, Section 5 presents some considerations and future work as conclusions.

2 Basic Concepts

In the following paragraphs we introduce the basic concepts on which our approach relies. We have restricted ourselves to term pairs in order to simplify the search of cluster labels. In the future work, we will extend this framework to consider unigrams and n-grams as cluster descriptors.

Let $\mathcal{C} = \{d_1, \dots, d_n\}$ be a collection of text documents, where each document $d_i \in \mathcal{C}$ is represented by a bag of terms¹. Let also $\mathcal{T}$ be the set of all terms that occur with a predefined minimum frequency in at least one document in $\mathcal{C}$, and $\mathcal{P}$ be the set of all pairs of terms in $\mathcal{T}$ that co-occur in documents of $\mathcal{C}$.

For a given $\pi \subseteq \mathcal{P}$, we denote by $\mathcal{C}|_\pi$ the *support set* of π in $\mathcal{C}$, i.e. the set of documents in $\mathcal{C}$ that contain both terms of π .

We define the probability of generating a pair of terms $\{t_i, t_j\} \in \mathcal{P}$ from $\mathcal{C}$ as:

$$P(\{t_i, t_j\}|\mathcal{C}) = \sum_{d \in \mathcal{C}} P(t_i|d)P(t_j|d)P(d|\mathcal{C})$$

where $P(d|\mathcal{C})$ is the probability of selecting document d from among all documents in $\mathcal{C}$, and $P(t_i|d)$ and $P(t_j|d)$ represent the probability of generating terms t_i and t_j from d respectively. As we consider each document in $\mathcal{C}$ to be equally probable, we estimate $P(d|\mathcal{C})$ as $1/|\mathcal{C}|$ for all $d \in \mathcal{C}$. The probability of generating a term t from a document d is estimated as the fraction $\frac{TF(t,d)}{\sum_{t' \in d} TF(t',d)}$, where $TF(t,d)$ is the number of occurrences of term t in d .

Given a set of documents $D \subseteq \mathcal{C}$ and a similarity function $s : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$, the β -similarity graph of D is an undirected graph whose vertices are the documents of D , and there is an edge between documents d_i and d_j if they are β -similar. Two documents d_i and d_j are β -similar if $s(d_i, d_j) \geq \beta$, where β is a user-defined parameter.

Entropy is often used as a characterization of the information content of a data source. For example, it has been used as measure for feature selection, lossless data compression methods, or for evaluating the quality of cluster partitions. In this paper, we use entropy for assessing the homogeneity of the contents referred by a set of documents. In order to measure the homogeneity of a collection, entropy is usually applied over the vocabulary of the collection at hand. However, it gives us little information about the number of topics a cluster is covering. For this reason, we propose an alternative way to estimate the homogeneity of a set of documents by analyzing the number of connected components it contains.

We define a set of documents D to be *homogeneous in contents* if the “pure” entropy of the partition induced by the connected components of the β -similarity graph of D is less than 1. The pure entropy of a partition $\Theta = \{\Theta_1, \dots, \Theta_k\}$ is calculated as follows:

---

¹ Terms correspond to meaningful word lemmas or phrases (e.g. named entities) directly extracted from texts. Stop words are disregarded.

$$H(\Theta) = - \sum_{i=1}^k P(\Theta_i|\Theta) \log_2 P(\Theta_i|\Theta)$$

where $P(\Theta_i|\Theta)$ can be estimated as $\frac{|\Theta_i|}{\sum_{j=1}^k |\Theta_j|}$.

As entropy expresses the average number of bits required to encode some information, this definition stems from the following fact: if less than one bit is needed to encode a partition, then it includes a predominant component representing the contents of the data set. We will call *core of the document set D* , denoted as $core(D)$, to this predominant component.

Let $\pi \in \mathcal{P}$ be a pair of terms whose support set $\mathcal{C}|\pi$ is homogeneous in contents. Let also $core(\mathcal{C}|\pi)$ be the core of this support set. We define the *set of relevant documents for the contents labeled by π* as:

$$Rel(\mathcal{C}|\pi) = core(\mathcal{C}|\pi) \cup \{d \in \mathcal{C} \mid \exists d' \in core(\mathcal{C}|\pi) [s(d, d') = \max_{d'' \in \mathcal{C}} s(d, d'')] \geq \beta]\} \quad (1)$$

That is, we consider as relevant documents for the contents labeled by a pair of terms all documents in the core of its support set, together with those documents in the collection $\mathcal{C}$ whose most β -similar document belongs to the core.

3 Our Approach

As previously mentioned, the goal of our proposal is to discover the topics from textual contents of a document collection, and simultaneously provide a short descriptive label for each one. We assume that no prior knowledge about the collection exists. Therefore, no topic samples are available and also the number of topics is unknown. As a consequence, all the processes involved in the proposed method (i.e. topic discovery and labeling) are completely unsupervised.

Given a text collection $\mathcal{C} = \{d_1, \dots, d_n\}$, our method obtains a clustering $\{(\ell_1, G_1), \dots, (\ell_m, G_m)\}$, where each cluster G_i represents a collection topic and ℓ_i is its label. Assuming that each cluster label can be selected from the set of term pairs $\mathcal{P}$, the clustering method relies on the probability of generating such labels from the collection to guide the search for a “good” partition of the data. A label is considered for producing a cluster if its support set is homogeneous in contents (i.e. it does not merge topics). The cluster is then defined as the set of relevant documents for the contents identified by the label. As previously mentioned, not all documents relevant to a label must have exactly such a label but a set of semantically related terms. For this reason, we also include in the cluster associated to a description all the documents whose most β -similar document is included in it (see Formula 1). Notice that the cluster keeps being homogeneous after adding these β -similar documents.

Thus, starting from the most probable pair of terms generated from the collection, its support set is built. If this set is homogeneous in contents, a cluster consisting of the set of relevant documents for the contents labeled by the pair is created. The pair becomes the cluster’s short descriptive label. In case that the

support set is not homogeneous in contents, the pair of terms is discarded. Once a cluster has been built, its documents are removed from the collection. Then, a new most probable term pair is selected from the remaining collection. This process is repeated until either the collection is empty or no more relevant pairs exist. Finally, if there are documents not clustered yet, a singleton is created for each one considering its most frequent pair of terms as its label. The general steps of our proposal are shown in Algorithm 1.

Algorithm 1. A clustering algorithm for obtaining labeled document clusters

---

Input: A text collection $\mathcal{C} = \{d_1, \dots, d_n\}$.

Output: A clustering $\mathcal{G} = \{(\ell_1, G_1), \dots, (\ell_m, G_m)\}$ of $\mathcal{C}$.

Build the set of term pairs $\mathcal{P}$

$\mathcal{G} = \emptyset$

$\mathcal{Y} = \mathcal{C}$

$\Pi = \operatorname{argmax}_{\{t_i, t_j\} \in \mathcal{P}} P(\{t_i, t_j\} | \mathcal{Y}) > 0$

while $\mathcal{Y} \neq \emptyset$ and $\Pi \neq \emptyset$ **do**

for all $\{t_i, t_j\}$ in Π **do**

if $\mathcal{Y}|_{\{t_i, t_j\}}$ is homogeneous in contents **then**

$\ell = \{t_i, t_j\}$

$G_\ell = \operatorname{Rel}(\mathcal{Y}|_{\{t_i, t_j\}})$

$\mathcal{G} = \mathcal{G} \cup \{(\ell, G_\ell)\}$

end if

$\mathcal{P} = \mathcal{P} \setminus \{\{t_i, t_j\}\}$

end for

$\mathcal{Y} = \mathcal{Y} \setminus \bigcup_{\ell} G_\ell$

$\Pi = \operatorname{argmax}_{\{t_i, t_j\} \in \mathcal{P}} P(\{t_i, t_j\} | \mathcal{Y}) > 0$

end while

if $\mathcal{Y} \neq \emptyset$ **then**

$\mathcal{G} = \mathcal{G} \cup \{(\{t_i, t_j\}, \{d\}) \mid d \in \mathcal{Y} \wedge t_i \text{ and } t_j \text{ are the most frequent terms in } d\}$

end if

return $\mathcal{G}$

---

4 Evaluation

For evaluating our approach we use two benchmark text collections. They are different in terms of number of topics, topic sizes, number of dimensions and document distribution. The first one consists of 8042 English texts from the first 100 topics of TDT2 dataset, version 4.0². The second is the AFP collection from TREC-5 conference³ that includes 695 Spanish articles manually classified into 25 topics. Each topic is provided with a title and a brief description in both collections.

---

² <http://www.nist.gov/speech/tests/tdt.html>

³ <http://trec.nist.gov>

In our experiments, documents are represented using the traditional Vector Space Model. Document terms are statistically weighted using SMART *ltc* [7]. The cosine similarity function is used to compare documents.

There are several measures to evaluate the quality of clustering algorithms. We adopt two traditional measures that compare the system-generated clusters with the manually labeled topics, namely: micro- and macro-averaged F1. Both measures combines the precision and recall factors. Whereas micro-averaging gives equal weight to every document, macro-averaging gives equal weight to each topic.

For each test collection, we consider a baseline directly built from its manual topics. The baseline consists of the document clusters that correspond to the support sets of the most probable term pairs of each manual topic.

We also consider two versions of our method in order to evaluate the impact of: (1) using only labels whose support sets are homogeneous in contents, and (2) adding those documents whose most β -similar document is included in the core of the support set of the labels (see Formula 1). For the first version, we disregard the homogeneity constraint imposed to the most probable pair of terms, and therefore clusters consist of the support sets of the pairs. Note that this version does not depend on the β parameter. For the second, we consider only the core of the probable term pairs for creating the clusters, instead of using Formula 1.

Figure 1 shows the F1 results obtained for each collection varying the β parameter. Several observations can be made by analyzing these results. First, it can be appreciated that our proposal obtains very good results for both micro- and macro-averaged F1 measures in both collections (around 0.82 and 0.78 in AFP collection, and 0.86 and 0.83 in TDT2 corpus of micro- and macro-averaged

Fig. 1. Micro- and macro- averaged F1 values obtained for AFP and TDT2 collections

F1 respectively). Micro- and macro-averaged values are highly correlated, which suggests that our approach is not affected by unbalanced topics. That is, small topics can be identified as well as large ones.

We can also see that in all cases the proposed method clearly outperforms both the baseline and the first version of our approach mentioned above. This indicates that using only term pairs is not enough for discovering topics, even though the pairs are generated from the manual topics. Moreover, these results corroborate the positive impact of filtering out term pairs considering the homogeneity of their support sets.

Regarding our approach w.r.t. its second version, it can be shown using the Wilcoxon signed-rank test [8] that there is a statistically significant improvement in the results when the β -similar documents are added to the core of the support set of the labels. In these collections the obtained p -values for micro- and macro-averaged F1 were $p \leq 1.15e-11$ and $p \leq 2.73e-6$, respectively.

In Table 1, we show the labels and F1 values obtained for some TDT2 topics. It can be seen that our method not only properly identifies the topics but also provides meaningful labels for each one. Notice the close correspondence between the obtained labels and the topic titles (as provided by the LDC annotators).

Table 1. Labels and F1 values obtained for some TDT2 topics

Topic title	Topic size	Cluster label	F1
McVeigh's navy Dismissal & Fight	11	{ <i>McVeigh, navy</i> }	0.75
Fossett's Ballon Ride	15	{ <i>ballon, world</i> }	1.00
Casey Martin Sues PGA	56	{ <i>Casey_Martin, cart</i> }	0.77
Mountain Hikers Lost	7	{ <i>avalanche, french</i> }	1.00
Pope visits Cuba	151	{ <i>pope, Cuba</i> }	0.71
Current Conflict with Iraq	1486	{ <i>Iraq, weapon</i> }	0.87
Bombing AL Clinic	99	{ <i>bombing, clinic</i> }	0.97
Cable Car Crash	110	{ <i>cable, car</i> }	0.89
China Airlines Crash	36	{ <i>plane, crash</i> }	1.00
Tornado in Florida	53	{ <i>tornado, Florida</i> }	0.88
Oprah Lawsuit	70	{ <i>winfrey, show</i> }	0.87

Finally, we compare the results obtained by our approach in the whole TDT2 collection to those reported in [3]. In that work it has been demonstrated that the use of a hierarchy of clusters improves notably the effectiveness of the system, as it allows the detection of topics with very different sizes. Despite our approach is not hierarchical, we obtain similar results (3.9% worst). This also confirms that our approach is not sensitive to the topic sizes. Other methods based on frequent sets [4,5,6] can only detect medium and large topics, because calculating frequent sets with low support use to lead a combinatorial explosion. Unfortunately, these methods cannot be compared with ours as they do not provide the actual set of topics over which they were evaluated.

5 Conclusions

In this paper, a new clustering algorithm has been introduced. It is aimed at obtaining labeled document clusters that accurately identify the topics of a text collection. The major novelty of the proposal is the combination of both the generation of probable term pairs and the estimation of the topic homogeneity associated to them for simultaneously determining both clusters and its labels. The experiments carried out over the TDT2 English corpus and AFP Spanish collection of TREC-5 validate the usefulness of the proposed method for both topic discovery and labeling.

We believe that the accuracy of our method could be improved by using a post processing phase in which tightly related clusters are combined and cluster labels are enriched. Future work also includes both estimating the parameter β and extending our approach in order to obtain topic descriptions in the form of abstract summaries.

References

1. Cutting, D.R., Pedersen, J.O., Karger, D.R., Tukey, J.W.: Scatter/Gather: a cluster-based approach to browsing large document collections. In: Belkin, N.J., Ingwersen, P., Pejtersen, A.M. (eds.) *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 318–329. ACM, Denmark (1992)
2. Zamir, O., Etzioni, O., Madani, O., Karp, R.M.: Fast and intuitive clustering of web documents. In: Heckerman, D., Mannila, H., Pregibon, D. (eds.) *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD 1997)*, pp. 287–290. AAAI Press, California (1997)
3. Pons-Porrata, A., Berlanga-Llavori, R., Ruiz-Shulcloper, J.: Topic discovery based on text mining techniques. *Inf. Process. Manage.* 43(3), 752–768 (2007)
4. Beil, F., Martin, E., Xu, X.: Frequent Term-Based Text Clustering. In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 436–442. ACM, Canada (2002)
5. Fung, B.C.M., Wang, K., Martin, E.: Hierarchical Document Clustering Using Frequent Itemsets. In: Barbará, D., Kamath, C. (eds.) *Proceedings of the Third SIAM International Conference on Data Mining*, pp. 59–70. SIAM, California (2003)
6. Li, Y., Chung, S.M., Holt, J.D.: Text Document Clustering Based on Frequent Word Meaning Sequences. *Data Knowl. Eng.* 64(1), 381–404 (2008)
7. Buckley, C., Salton, G., Allan, J., Singhal, A.: Automatic Query Expansion Using SMART: TREC-3. In: *Proceedings of the Third TREC Conference*, Washington, pp. 69–80 (1994)
8. Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics* 1(6), 80–83 (1945)