

# Highlighting latent structures in texts

Michèle Jardino

LIMSI-CNRS  
BP 133, 91403 Orsay cedex, France

## Abstract

We have developed an original learning method in order to extract latent structures in raw texts. The induced structure is a data-driven tree which can be unbalanced. It has been obtained from successive partitions of the texts in clusters, with an incremental number of classes ranging from 2 to  $K$ ; each quasi-optimal partition has been performed with an adaptation of the k-means clustering. The paths of the texts in the successive partitions are the edges of an oriented graph whose nodes are the clusters. The study of the paths shows that some of the clusters remain identical in the successive partitions so that a tree can be extracted from the graph, by merging nodes and clipping edges. A corpus of 1,100 touring information leaflets has been used to illustrate this method.

## Problem

Giving a set of raw texts, how to group them in valuable clusters without the a priori knowledge of the number of clusters?

Usually, two types of unsupervised clustering are considered to automatically group data: the hierarchical clustering (top-down or bottom-up) and the k-means clustering. Today, the determination of the number of clusters remains an open question (Duda and Hart 1973, Kaufman and Rousseeuw 1990, Lebart et al. 1995). Here, we propose a method to learn this number from the data and we will show how this method induces the discovery of latent structures in the texts. It is worthwhile to notice that this method can be applied to any type of data.

## Proposition

We make a set of quasi-optimal partitions of the texts. Each partition corresponds to a specific number of clusters, ranging from 2 to  $K$ . We obtain  $K-1$  partitions and we search the clusters which remain stable among the successive partitions.

We build up a lattice with the paths of the texts in the partitions and if there exists a structure in the set of texts, a tree can be extracted from the lattice by merging clusters in the partitions and clipping paths between identical clusters.

## Text and cluster representation

We consider texts as bags of words and represent them by vectors whose components are the relative frequencies of the words in the texts. In statistical words it corresponds to the profiles of the texts. We write  $f_i^j$  the relative frequency of the word  $j$  in the text  $i$  and  $l_i$  the length of the text  $i$ .

The corpus is a set of 1,100 touring information leaflets whose the descriptive part has been selected. In order to improve information retrieval in the leaflets, we have segmented each leaflet into several chunks. About 4,700 chunks have been obtained. It constitutes our learning text set. It corresponds to about 120,000 words with a vocabulary size equal to 6,300 words.

Texts are grouped in clusters, represented by their centroids. Assuming that  $f_c^j$  is the relative frequency of the word  $j$  in the cluster  $c$  and that  $l_c$  is the sum of the lengths of the texts grouped in the cluster, the centroid of this cluster is defined as:

$$l_c \times f_c^j = \sum_{i \in c} l_i \times f_i^j$$

### Unsupervised clustering: k-means like algorithm

K-means algorithm is an iterative method in which items are grouped in clusters, represented by their centroids. At each step, one item is moved from one cluster to another one, in order to maximise the sum of the distances between the centroids. We have modified this algorithm in the following ways:

- 1- instead of the geometrical measure (euclidean distance) we have used an entropy-based measure. The entropy of the texts grouped in  $k$  clusters is:

$$H(k) = -(1/f) \sum_{c=1}^k \sum_{j=1}^V f_c^j \times \log(f_c^j / l_c)$$

where  $V$  is the size of the vocabulary and  $f$  the total number of words in the corpus.

When  $k$  is given,  $H(k)$  depends on how the texts are distributed in the clusters. With the clustering algorithm, we search the partition which gives the lowest value of  $H(k)$  (Jardino 2000).

- 2- instead of the greedy search of an optimal solution at each iterative step, a new clustering is accepted when the new entropy is simply lower than the entropy of the preceding step. The movings are randomly performed.

With this method, we obtained quasi-optimal partitions which almost do not depend on the initial conditions.

### Successive partitions from 2 to K clusters

Firstly, an optimal partition of the text in two clusters is performed with the k-means like algorithm. At the beginning of this clustering, the texts are grouped into one cluster. Then, the next clustering processes are initialized with the results of the preceding clustering processes : the clustering of the chunks in  $k$  clusters is initialized with the results of the clustering in  $k-1$  clusters. A small bias is created by this procedure because the partitions are only quasi-optimal but this makes the paths between the stable clusters more obvious. The optimal partitions in 2, ...,  $k$ , ...,  $K$  clusters are respectively labelled  $P_2$ , ...,  $P_k$ , ...,  $P_K$ .

Nine partitions of the corpus have been performed, from 2 to 10 clusters. In the table 1 are gathered some values related to the corpus and the clustering.

Number of texts	4 700 chunks
Vocabulary size	6 300 words
Total number of words	120 000 words
Maximum entropy	377 words
Minimum entropy	25 words
Entropy, 2 clusters	295 words
Processing time, 2 clusters	2 s
Entropy, 10 clusters	162 words
Processing time, 10 clusters	28 s

Table 1: Partitions of the chunks

The value associated to the entropy is given in terms of the number of words, it is the exponential value of the entropy (Jelinek, 1988).

### Tree search

We draw an oriented graph whose nodes are the clusters and whose edges are the paths of the chunks in the successive partitions. There are  $k$  layers in the lattice, and a maximum of  $k!$  edges is possible between the clusters in the successive layers. In our example, only 295 edges exist while  $10!$  edges are possible. This elementary measure shows that there are favorite paths between the layers. The ten most frequent paths represent 75% of the corpus. On the figure 1 is shown the corresponding graph.

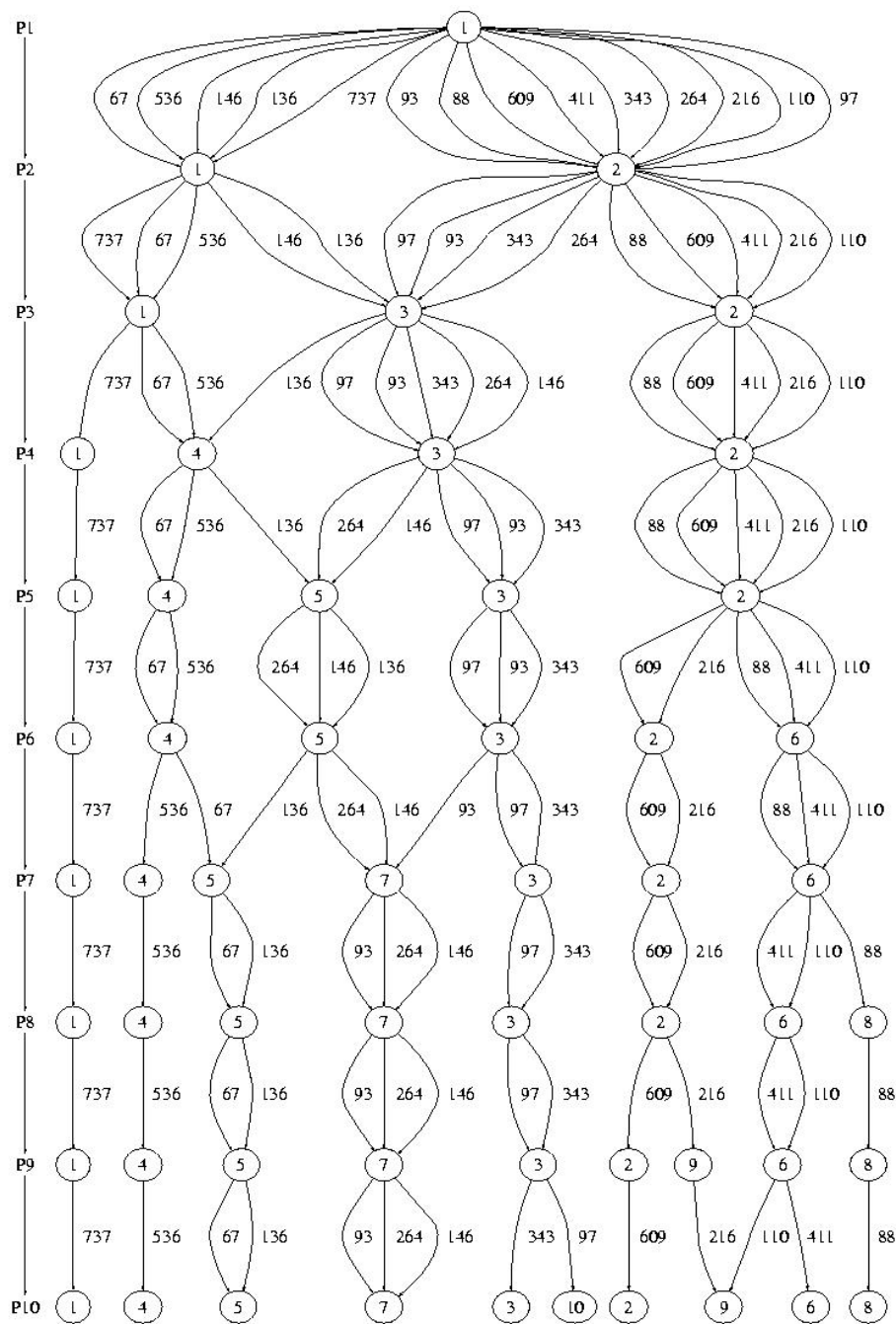


Figure 1 : Lattice of the paths of the chunks in the partitions from 2 to 10 clusters.

The numbers near the edges are the number of chunks which belong to the clusters connected by the edges. This figure shows the behaviour of the clusters in the successive partitions. We observe that several clusters remain identical. It is then possible to construct a tree from the graph.

We observe that the partition 8 is stable, so we cut the lattice at this level and we consider the bottom-up paths. We clip isolated branches and merge clusters which are stemming from more than one cluster in the partition of the lower order. The merging of the clusters 4, 5, 7 and 3 of P7 corresponds to the merging of the clusters 4, 5 and 3 of P6 and P5 and to the merging of the clusters 4 and 3 of P4: one single node is associated to these clusters in the new branching, the label of this cluster is the labels of the clusters in the partition close to P1 in the lattice, that is to say "4,3 (P4)". In the same branch, the clusters 1 and 3 of P3 are merged in the cluster "1,3 (P3)". The clusters 6 and 8 of P8 are merged in the cluster 6 of P7 and P6, its label is, according to the preceding rule, 6 (P6). The clusters 2 and 6 of P7 and P6 are merged in the cluster 2 of P5 which is the same in P4 and P3, its label is "2 (P3)". The cluster 1 of P8 is merged with the cluster 3 of P3, its label is "1 (P4)", because this cluster is isolated from P4.

This gives the next unbalanced tree where the "new clusters" are labelled with the labels of the clusters in the initial graph and with the labels of the associated partitions.

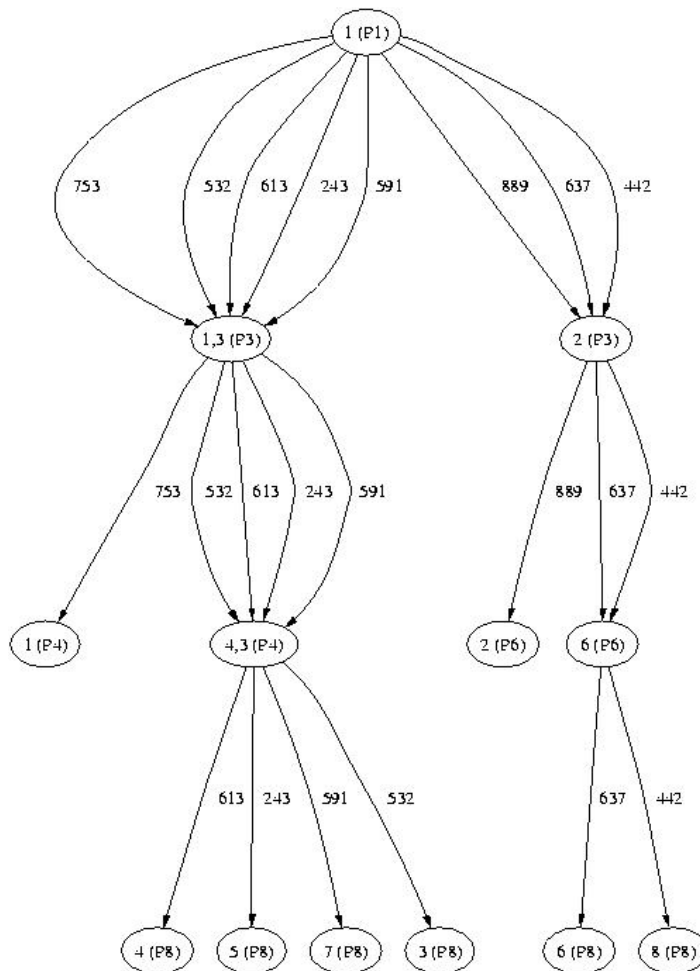


Figure 2 : Tree extracted from the lattice

## Stability criteria

We bring to the fore two criteria for the automatic detection of stable clusters. The first one is based on the paths between the partitions, the second one on the cluster entropies.

### *First criterion : paths between successive partitions*

One cluster of the partition  $P_k$  is said stable if its items (texts) are issued from only one cluster of the partition  $P_{k-1}$ . When all the clusters of the partition  $P_k$  are stable, there are  $k$  ways between  $P_k$  and  $P_{k-1}$ . In the opposite case, when all the clusters are unstable, there are  $k*k-1$  ways between these partitions. In conclusion, the number of observed paths between successive partitions seems to be a valuable indicator of the stability of the clusters, it varies between  $k$  and  $k*k-1$ . On the figure 3 are drawn the variations of the number of paths observed between successive partitions when  $k$  varies from 2 to 15.

We observe that this number remains constant between  $P_7$  and  $P_8$ . It is consistent with the lattice (figure 1).

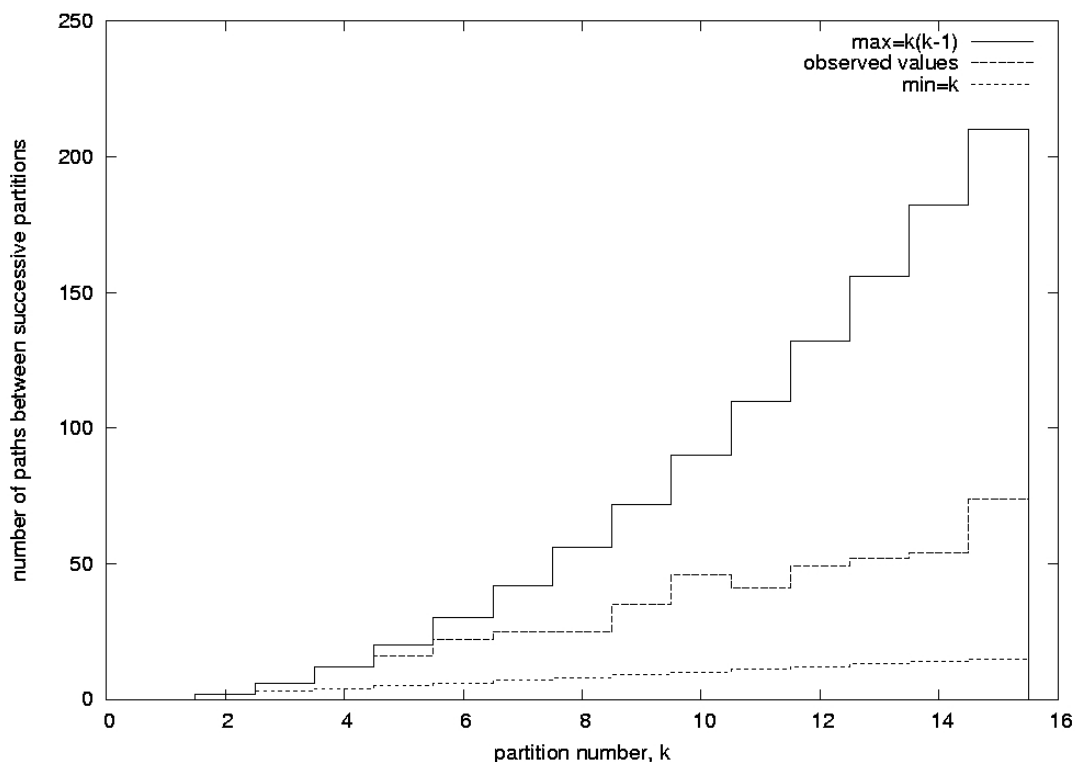


Figure 3: Number of paths, observed between the successive partitions of the chunks and compared with theoretical minima and maxima.

### *Second criterion : cluster entropy*

We have drawn on the figure 4 the variation of the entropy in four typical cases. The entropy of the cluster 1 decreases rapidly in three steps (from  $P_2$  to  $P_4$ ) and then remains stable. The entropy of the cluster 2 varies strongly between  $P_1$  and  $P_3$ , then varies slowly between  $P_3$  and  $P_5$ , strongly from  $P_5$  to  $P_6$ , then the variations are weak. We can deduce that this cluster is

stable in P3, P4 and P5. It is split at P6 and is stable in P7 and P8. There is a new split in P9. The cluster 2 seems to be an homogeneous cluster which evolves by splitting. The two important variations of entropy of the cluster 4 are related respectively to the split of the cluster 4 (P4) in the clusters 4(P5) and 5(P5), and to the split of the cluster 4(P6) in the clusters 4(P7) and 5(P7). The entropy variations of the cluster 10 show that increases are possible at a local level.

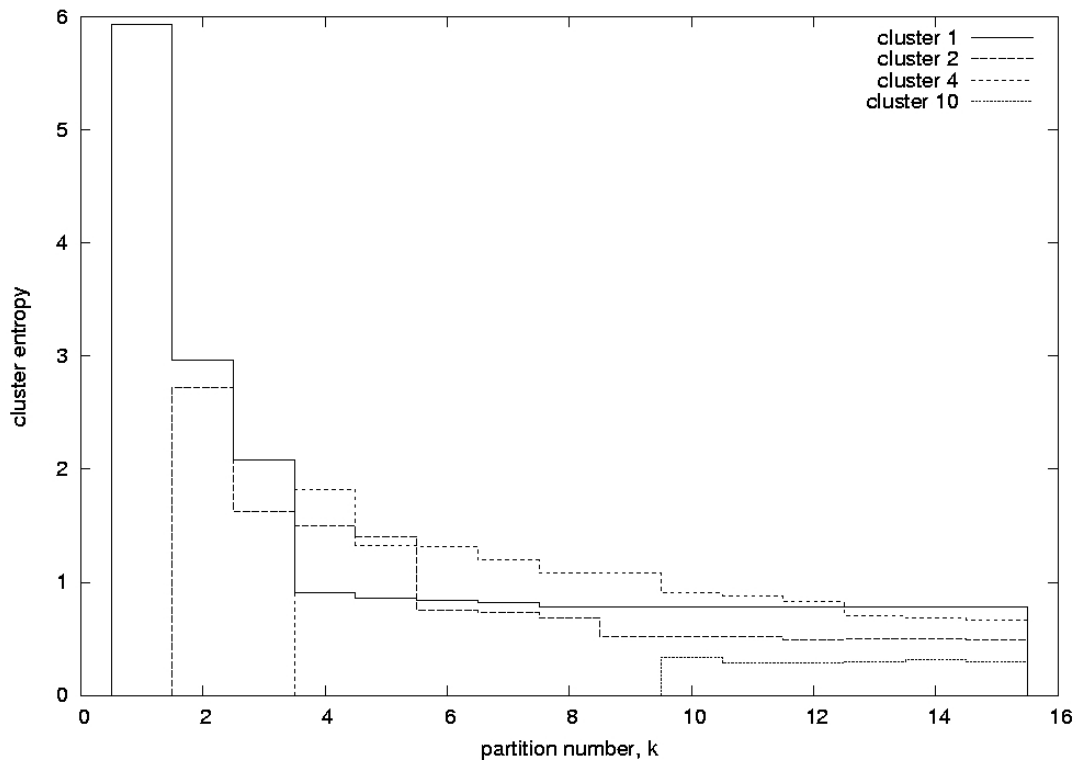


Figure 4: Entropies of the clusters 1, 2, 4 and 10 in the successive partitions

We observe a variation of entropy strongly related to the paths in the lattice : large variations correspond to splits and stabilities correspond to either stable clusters or merged clusters. Actually, a null variation of the entropy of one cluster does not imply that this cluster is stable. This is why this criterion which is largely used to determine an optimal value of k (Duda and Hart) does not give satisfactory results. It can be only considered as an help to find the latent structure.

### Preliminary assessment

An XML file has been created according to the tree structure of the data (figure 2). With Xpath, a language for addressing parts of an XML document, a few queries have been used to test the retrieval performances of the structured documents. For each query, the hierarchical representation improves the precision of the returned documents. For example, the question “resort overlooking a lagoon” returns 19 leaflets within our framework whereas 48 leaflets are given by the non-hierarchical data. The 29 additional leaflets correspond to “lagoon swimming pool” (Folch 2004).

## Conclusion

We have shown how to extract a latent structure in textual data : first, successive partitions of the texts are performed with an unsupervised clustering algorithm, then stable clusters are detected. These stable clusters serve as anchors to create a tree which is the image of the latent structure of the data. A first experiment has shown that this structure improves information retrieval over the leaflets.

Two problems have to be solved :

- 1- how to automatically find at which layer the clusters remain stable, as today a part of our method is visual
- 2- how to label the clusters with sensible tags.

The solution of the first point seems easy, the number of observed paths between the clusters in successive partitions seems a good criterion. A possible clue to the second point is the search of the discriminant words for each cluster.

## Références

- Celeux G., Diday E., Govaert G., Lechevallier Y., Ralambondrainy H. (1989). *Classification automatique des données*. Dunod.
- Cover T. and Thomas J. (1991). *Elements of Information Theory*. Wiley & sons.
- Duda R.O. and Hart P.E. (1973). *Pattern classification and Scene Analysis*. Wiley & sons.
- Folch H., Habert B., Jardino M., Pernelle N., Rousset M.C., Termier A. Submitted to LREC2004.
- Jardino M. (2000). Unsupervised non-hierarchical entropy-based clustering. In Kiers H.A.L., Rasson J.-P., Groenen P.J.F., Schader M. editors, *Data Analysis, Classification and Related Methods*, p.29. Springer.
- Jelinek F. (1988). *Statistical Methods for Speech recognition*. MIT Press, Cambridge, Massachusetts.
- Kaufman L., Rousseeuw P.J. (1990). *Finding groups in data*. Wiley & sons.