

# A simple mixture model for unsupervised text categorisation

F. Clérot<sup>1</sup>, F. Fessant<sup>1</sup>, O. Collin<sup>1</sup>, O. Cappé<sup>2</sup> & E. Moulines<sup>2</sup>

<sup>1</sup>France Télécom R&D, 2 avenue Pierre Marzin, 22300 Lannion, France

<sup>2</sup>Télécom Paris, 46 rue Barrault, 75013 Paris, France

## Abstract

Automatically segmenting text corpora into thematically related groups is a complex exploratory analysis problem. In this article, we outline our multi-stage exploratory analysis process and investigate the performance of a simple statistical model. After a description of this model and of its fitting procedure, we illustrate its performance on the segmentation of a corpus of CKM-related texts in English.

*Keywords: text mining, exploratory analysis, clustering, mixture model.*

## 1 Introduction

Clustering is a key tool in exploratory data analysis; segmenting the data into homogeneous groups leads to a more synthetic understanding of the data, allows to build powerful visualisations and is often the first step towards more specific analysis such as supervised classification.

Although less standard in analysis of text data, clustering has recently received a lot of attention. The goal is to bring to text data analysis the same benefits as above.

There are however significant differences between text data analysis and numerical data analysis: for numerical data analysis, the cluster "homogeneity" is judged from a metric in data space; when dealing with text data, it is clearly implicit that "homogeneity" means "topical homogeneity", a notion which is more difficult to measure.

In this article, the purpose of text clustering is to build a topical segmentation of a corpus. Because of the difficulty of defining a priori a topical homogeneity measure, the text clustering analysis must be considered as a part of



a complete exploratory analysis process and therefore must allow the analyst to interactively monitor the process so as to select the most relevant features. Powerful visualisation will make this interaction more natural.

Such a semi-automatic construction of a topical segmentation is a very helpful tool for automatic technological survey and is also expected to be a very useful step towards ontology learning and knowledge management.

The mixture model is introduced in section 2; its use inside a complete exploratory data analysis is described in section 3. Section 4 gives the details of the fitting of the model. The results of a case study are given in section 5 as an illustration of the performance of the approach.

## 2 A simple mixture model

Text documents have to be brought to a numerical representation so that statistical methods can be applied. We use the simplest representation, reducing the document to a "bag of words": a document is represented by a vector of the number of occurrences in this document of "words" taken from a given dictionary.

A corpus of texts is therefore represented as a set of points in a high dimensional vector space (there are as many dimensions as words in the dictionary); our clustering approach relies on the estimation on the density of this set of points.

We introduce the following notations: "words" are indexed by  $w = 1, \dots, W$ ; documents are indexed by  $d = 1, \dots, D$ . The corpus of documents is therefore represented by a matrix of counts  $\left( c_d^w \right)_{d=1 \dots D}^{w=1 \dots W}$ .

Our target topical segmentation has the simplest possible structure,  $T$  mutually exclusive topics, indexed by  $t = 1, \dots, T$ .

There are two main hypotheses underlying the model:

- a document  $d$  is *a priori* associated to one and only one topic  $t$ ;
- conditionally to a topic  $t$ , the word counts are independently distributed.

Under these hypotheses, we can express the probability of a document  $d$  as a mixture of laws:

$$\Pr \left[ \left( c_d^w \right)_{w=1 \dots W} \right] = \sum_{t=1}^T \Pr(t) \prod_{w=1}^W \Pr \left( c_d^w | t \right) \quad (1)$$

where  $\Pr(t)$  is the probability of topic  $t$  and  $\Pr \left( c_d^w | t \right)$  is the probability of the count  $c_d^w$  of word  $w$  in document  $d$ , conditionally to topic  $t$ .



The specific form of these probability laws is chosen as follows:

- the distribution of the topics follows a multinomial law  $Mult(1; p_1, \dots, p_T)$ ;
- the counts of word  $w$  conditionally to the topic  $t$  follow a Poisson law with parameter  $\lambda^w(t)$ .

The choice of a Poisson law to represent the word counts is consistent with the observation that count matrices are generally very sparse and with our choice of having a very simple model.

### 3 Exploratory data analysis vs. statistical analysis

The mixture model defined in eqn (1) has already been introduced and discussed in the literature, although with a different choice for the count law; see Nigam et al [1] and Nigam et al [2]. The performance evaluation in Blei et al [3] shows that the first hypothesis above may be much too restrictive to be applied directly to a corpus of documents: obviously, a document may span many topics, especially if it is a long one. The authors therefore propose a more complex model which *a priori* allows a document to belong to any theme (Latent Dirichlet Allocation, LDA). As shown in Girolami and Kaban [4], this model appears as a generalisation of pLSI (probabilistic Latent Semantic Indexing, Hoffman [5]). In both cases, the authors aim at capturing the statistical structure of the corpus in only one step.

Our approach is quite different: our exploratory data analysis process involves successive steps of increasing generality. Instead of trying to capture the statistical structure of a corpus of documents with very different lengths (as is often the case) in just one step, we rather try to successively identify an "atomic" level in the documents where a simple model (as the mixture defined in eqn (1)) can be applied, build the clustering at this level, give an interpretation to the clusters, consider the documents as sets of their atomic parts and build a document representation according to the clusters of their atomic parts.

We have already used such a multi-stage exploratory data analysis process in various domains (see Cl  rot and Fessant [6] for instance). The interpretation step at the end of each stage necessarily involves the analyst (and possibly experts of the domain) and allows to incrementally extract knowledge from the analysis results. As we shall see below, this crucial interpretation step is often not trivial, even with such a simple model where a topic is simply a combination average occurrence of words. Complex models which allow a word to belong to several topics inside the same document may prove even more difficult to interpret in realistic situations, although they can be more faithful to the distribution of the data.

We believe that our multi-stage approach represents a good practical compromise between the quality of the density estimation and the interpretability of the model since each level gradually increases the complexity, leveraging the



knowledge extracted from the underlying levels to build a relevant representation.

A last point is worth mentioning: as our target topical segmentation is made of  $T$  mutually exclusive topics, we expect the *a posteriori* document classification to unambiguously assign most documents to one topic only. It is therefore important to *a priori* enforce this property in the model.

In the case of text documents, an obvious "atomic" level is the paragraph level and constraining a paragraph to belong to only one topic seems very natural, so that a model as in eqn (1) will be quite adapted at this level.

## 4 Model fitting

The parameters defining the model are the number of topics  $T$ , the probabilities of the topics,  $(p_t)_{t=1 \dots T}$  and the Poisson parameters for the counts of words conditionally to the topics  $\left[\lambda^w(t)\right]_{t=1 \dots T}^{w=1 \dots W}$ . The number of topics is given by the analyst; only the topic probabilities and the Poisson parameters are estimated.

The mixture model of eqn (1) is fitted by maximizing the likelihood with EM (Titterton et al [7]). Priors on the distribution of the parameters are introduced:

- a Dirichlet law with parameters  $(\theta, \dots, \theta)$ , conjugate of the multinomial law
- a Gamma law with parameters  $(\alpha, \beta)$ , conjugate of the Poisson law.

The same values of the parameters are used for all laws. Using conjugate laws allows to give a simple analytic form to the EM re-estimation formulas (Denison et al [8]). We obtain:

$$p_t = \frac{(\theta - 1) + \sum_{d=1}^D \gamma_d(t)}{T(\theta - 1) + \sum_{t=1}^T \sum_{d=1}^D \gamma_d(t)} \quad (2)$$

$$\lambda^w(t) = \frac{(\alpha - 1) + \sum_{d=1}^D c_d^w \gamma_d(t)}{\beta + \sum_{d=1}^D \gamma_d(t)} \quad (3)$$

where we introduced



$$\gamma_d(t) = \frac{p_t \times \exp(-\sum_{w=1}^W \lambda^w(t)) \times \prod_{w=1}^W \lambda^w(t)^{c_d^w}}{\sum_{s=1}^T p_s \times \exp(-\sum_{w=1}^W \lambda^w(s)) \times \prod_{w=1}^W \lambda^w(s)^{c_d^w}} \quad (4)$$

which can be interpreted as the a posteriori probability for document  $d$  to belong to topic  $t$ .

The algorithm is initialized from a matrix  $\left[\gamma_d(t)\right]_{d=1 \dots D}^{t=1 \dots T}$  randomly drawn between 0 and 1 and properly normalised. Such initialisation leads to very stable results from one run to another.

As obvious from eqns (2) and (3), the parameters of the conjugate laws allow to influence the behaviour of the solution; choosing  $\theta$  smaller than 1 will favour a distribution of theme probabilities either large or very small. The results below are shown for  $\theta = 0.5$  ("Jeffrey's prior") but, in this case, we did not notice significant differences as compared to  $\theta = 1$  (uniform prior). For the Gamma law, we set its mean to the empirical mean of the counts in the corpus and chose a variance much larger than the empirical variance.

At the end of the estimation process, we obtain the values of the parameters of the model; with these parameters, eqn (4) allows to calculate the probability for any document to belong to a topic, given that this document is described as a count vector  $\left(c_d^w\right)_{w=1 \dots W}$ . Therefore, eqn (4) also allows to classify new documents, not only documents from the corpus.

As remarked above, the number of topics  $T$  must be set a priori by the analyst. It is possible to select this value by analysing the variation of the model likelihood as a function of  $T$  and choosing a value above which the likelihood does not increase significantly anymore as more topics are allowed.

## 5 Model performance on a case study

We present below the results of a case study on a corpus of CKM (Customer Knowledge Management)-related texts in English.

### 5.1 Corpus description and pre-processing

The automatic construction of a topical segmentation on such a restricted domain is a very difficult problem for traditional methods since the available resources are much too general to accurately resolve categories inside the domain. The "manual" acquisition of specialised resources is of course possible but takes a lot of time and often necessitates the collaboration of an expert of the domain who may not be easily available. Moreover, such manual acquisition clearly does not



scale for activities such as technological survey where the domains of interest may be very restricted and constantly changing.

The corpus is taken from a group of HTML documents of varying lengths whose tags have been "cleaned up", and which have been divided into words, phrases and paragraphs by means of a classical segmentation treatment. A labelling system based on syntactic analysis then projects the segments on a group of lemmas associated with well-defined syntactic categories. Preliminary tests have shown that grouping lemmas according to these sub-categories is not useful for our problem. Therefore we have retained a more general level of representation which corresponds to conventional categories (noun, verb, adjective etc.). The lemmas of proper nouns are the proper nouns themselves. English grammatical words as well as very frequently and very rarely-occurring words are ignored and we keep 1000 lemmas for the description of the documents. Unique identifiers are then associated to each lemma, document and paragraph, allowing the constitution of paragraph-lemma count matrix.

As indicated above, the results presented here concern the corpus of the paragraphs. Ignoring the paragraphs with less than 10 lemmas, the corpus finally contains 12489 different paragraphs, represented by a 1000-dimensional vector.

A first run of the algorithm revealed some groups which were clearly clustered around some idiosyncrasies of the corpus: as this corpus contains many extracts of Gartner Group or Datamonitor reports, we found clusters of paragraphs mainly characterised by the "*gartner*" or "*datamonitor*" lemmas. Although natural from the strict point of view of our algorithm, this behaviour is not satisfactory from an applicative point of view and we manually suppressed some lemmas leading to such clusters, ending up with a description on 973 lemmas.

## 5.2 Result visualisation and interpretation

We present the results obtained for  $T = 20$  topics. All the obtained topics have significant probabilities (see Table 1 below). We check that the a posteriori paragraph classification was consistent with the "one topic per document only" a priori:

- for 75% of the paragraphs, the "major topic" (the topic with the largest probability) has a probability greater than  $1/2$ , the 19 other topics sharing the other half;
- for 50% of the paragraphs, the major topic has a probability greater than 0.9, the 19 other topics sharing the remaining 0.1.

This shows that the a priori hypothesis of a single topic per paragraph is well followed by the a posteriori distribution and that the resulting clustering allows to categorize most documents into a single category with little ambiguity.

Interpretation of the categories is an important step in our analysis process; on such a specialised domain as CKM, advice from an expert of the domain is mandatory. Such experts are not easily available and this collaboration should be as productive as possible; it is greatly enhanced if they can share a single



"intuitive" vision of the domain. The visualisation of the results described below is a first step in this direction.

For lack of space, we only briefly sketch our visualisation technique below; it is inspired by Brandes and Cornelsen [9] and Brandes and Willhalm [10] to which we refer for motivation and additional details.

For each topic, we only keep the 30 words with the largest relative frequency inside the topic ("keywords") and we calculate the co-occurrence matrix of these words inside the topic  $K^t$  and on the whole corpus  $K^{corpus}$ .

We normalise the counts inside the topic to the counts on the whole corpus as  $\hat{K}_{i,j}^t = K_{i,j}^t / K_{i,j}^{corpus}$ ; this *symmetric* matrix expresses the coupling between keywords inside the topic  $t$  and we visualise this coupling by projecting the keywords on the first two non trivial eigenvectors of its Laplacian.

We normalise  $K^t$  line-wise to obtain  $(P_{i,j}^t)$ , the *asymmetric* matrix of the probabilities for word  $i$  to be in the same document as word  $j$  conditionally to the topic  $t$  and calculate the *authority* of each keyword from this matrix.

An example of such a visualisation is given on Figure 1; the vertex size indicates the authority of the associated keyword.

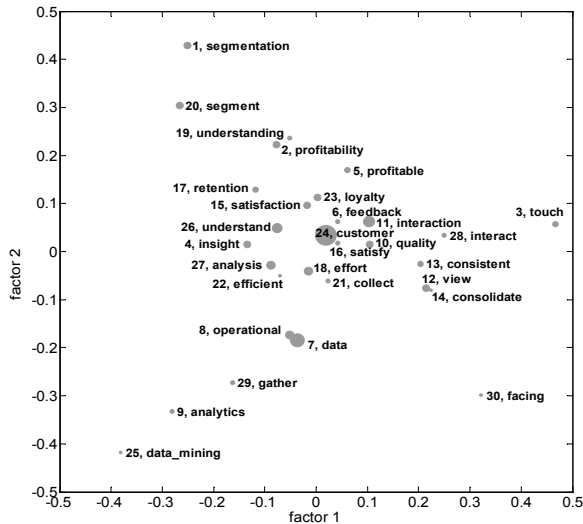


Figure 1: Visualisation of the first topic.

Figure 1 shows the projection for the dominant topic in the corpus; most of the keywords are CKM "buzzwords" and paragraphs in this topic mainly deal with a general introduction to the CKM domain. The central role in the graph is



played by "customer" which also happens to have the highest authority indicating that many other keywords have a strong probability to appear in the same document conditionally to the topic. The most prominent keyword in terms of relative frequency, "segmentation", only appears on the edge of the graph and has a rather poor authority : relationships between words (as measured from co-occurrences) are more informative than frequencies (which are the raw results of the analysis) for the interpretation of the topics.

Interestingly, the three clusters expanding around the centre deal with analytics, marketing and interaction, three well-identified dimensions of CKM. this visualisation indeed gives a very good "big picture" of the CKM domain.

Another visualisation is shown on Figure 2; it is easily seen that the topic is related to call-centres and, again, the three main clusters expanding around the centre have well-identified meanings, dealing respectively with CTI (Computer Telephony Integration), staff-related issues and services.

These two examples are illustrative of the strong interpretative power of our visualisation scheme; they allow the experts to rapidly and accurately interact with the results of the analysis. Moreover, such visualisations are very easily generated from the analysis raw results.

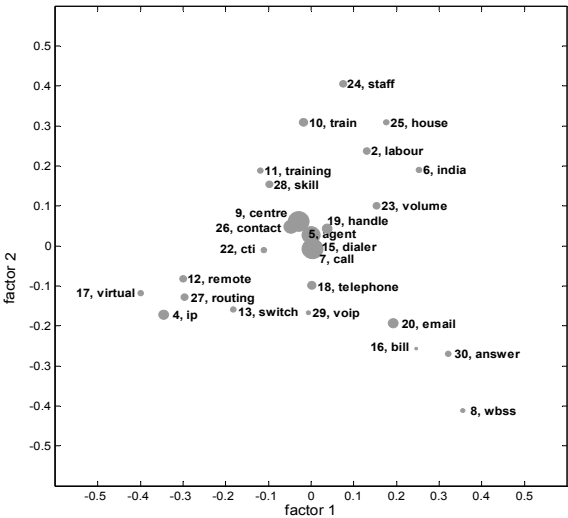


Figure 2: Visualisation of the topic "call centre".

Interpretation has however not been possible for all clusters; table 1 sums up the results: interpretations may be unambiguous, dubious or unclear; the figures in the table are the respective probabilities  $p_i$  of the topics (in percentages).

These results are quite satisfactory since the most prominent topics can be unambiguously interpreted and the topics with a clear interpretation gather more

than 60% of the paragraphs. Moreover, it can be seen that some small clusters can also be unambiguously interpreted; therefore, considering that the largest clusters only are relevant for the interpretation would waste valuable information.

Table 1: Interpretation of the clusters.

| <b>Interpretation</b>            | <b>Unambiguous</b> | <b>Dubious</b> | <b>Unclear</b> |
|----------------------------------|--------------------|----------------|----------------|
| General CKM introduction         | 8.5                |                |                |
| CKM project management           | 7.5                |                |                |
| On-line sales                    | 7.2                |                |                |
| Content personalisation          | 7.1                |                |                |
| Call-centres                     | 6.9                |                |                |
| Geographical comparisons         | 6.3                |                |                |
| Bank and insurance               | 6.3                |                |                |
| ?                                |                    |                | 5.0            |
| ?                                |                    |                | 4.7            |
| Natural language processing      | 4.6                |                |                |
| ?                                |                    |                | 4.2            |
| Partner relationship management  |                    | 4.0            |                |
| ?                                |                    |                | 3.9            |
| ?                                |                    |                | 3.9            |
| ?                                |                    |                | 3.8            |
| Mobility                         | 3.8                |                |                |
| ?                                |                    |                | 3.5            |
| Reuters idiosyncrasy             |                    | 3.1            |                |
| Automatic web site configuration |                    | 3.0            |                |
| Licence fees for CKM softwares   | 2.9                |                |                |

## 6 Conclusion

We have presented a simple statistical model and its use inside a complete exploratory data analysis process for the semi-automatic unsupervised categorisation of text data. Our step-by-step approach aims at striking a practical balance between the faithful representation of the data and the interpretative power of the resulting clustering.

We also described a simple visualisation technique which greatly simplifies the interpretation.

The performance of our simple model has been illustrated on a practical case study with a corpus on a very restricted domain; we obtained a very fine-grained topical segmentation where the most important topics can be unambiguously interpreted and the associated classifier unambiguously attributes most of the document to one topic only.



## References

- [1] Nigam K., Lafferty J., McCallum A., "Using maximum entropy for text classification", IJCAI, Workshop on Artificial Intelligence for Information Filtering, 61-67, 1999.
- [2] Nigam K., McCallum A., Thrun S., Mitchell T., "Text classification from labelled and unlabelled data using EM", *Machine Learning*, 39(2/3):103-134, 2000.
- [3] Blei D., Ng A., Jordan M., "Latent Dirichlet allocation", *Journal of Machine Learning Research*, 3(5):993-1022, 2003.
- [4] Girolami M., Kaban A., "On an equivalence between pLSI and LDA", in *Proceedings of the 26<sup>th</sup> Annual ACM Conference on Research and Development in Information Retrieval, SIGIR*, 433-434, 2003.
- [5] Hoffman T., "Probabilistic latent semantic indexing", in *Proceedings of the 22<sup>nd</sup> Annual ACM Conference on Research and Development in Information Retrieval*, 50-57, 1999.
- [6] Clérot F., Fessant F., "From IP port numbers to ADSL customer segmentation: knowledge aggregation and representation using Kohonen maps", *Data Mining IV (Rio de Janeiro)*, Eds Ebecken, Brebbia, Zanasi, WIT Press (2004).
- [7] Titterington D., Smith A., Makov U., "Statistical analysis of finite mixture distributions", Wiley (1985).
- [8] Denison D., Holmes C., Mallick B., Smith A., "Bayesian methods for nonlinear classification and regression", Wiley (2002).
- [9] Brandes U., Cornelsen S., "Visual ranking of link structures", *Journal of Graph Algorithms and Applications*, 7, 2, 181-201, 2003.
- [10] Brandes U., Willham T., "Visualization of Bibliographic Networks with a Reshaped Landscape Metaphor", preprint Nr. 170 from Universität Konstanz, 2002.

