INFERENCE FOR PROBABILISTIC UNSUPERVISED TEXT CLUSTERING

Loïs Rigouste, Olivier Cappé, François Yvon

Ecole Nationale Supérieure des Télécommunications (GET / CNRS UMR 5141)
46 rue Barrault, 75634 Paris Cedex 13, France
(e-mails: rigouste, cappe, yvon at enst.fr)

ABSTRACT

In this article, we investigate the use of a simple probabilistic model for unsupervised document clustering in large collections of texts. The model consists of a mixture of multinomial distributions over the word counts, each component corresponding to a different theme. The Expectation-Maximization (EM) algorithm is the basic tool used for inference.

After introducing the model and experimental framework (corpus and evaluation measures), we discuss the importance of initialization and illustrate the difficulty caused by the lack of supervision information. We propose some ideas to solve this problem, one of the most efficient method being based on vocabulary reduction, and finally compare those heuristics with other inference processes, such as Gibbs Sampling.

1. INTRODUCTION

Due to the wide availability of huge collections of text documents (news corpora, e-mails, web pages, scientific articles...), unsupervised clustering has emerged as an important text mining task. Several probabilistic models, performing a non-deterministic clustering of the data, such as Probabilistic Latent Semantic Analysis [1] or Latent Dirichlet Allocation [2], have been introduced for that purpose. In this contribution, we study the simpler model [3, 4] in which the corpus is represented by a mixture of multinomial distributions, each component corresponding to a different "theme". Dirichlet priors are set on the parameters and we use the Expectation-Maximization (EM) algorithm to obtain maximum *a posteriori* (MAP) estimates of the parameters.

To start with, we introduce the model and notations used throughout the paper. We then describe our evaluation framework and highlight, in first round of experiments, the importance of the initialization step in the EM algorithm. Looking for ways to overstep the limitations of EM by incremental learning, we present a hierarchical algorithm and other ideas based on variations in the size of the vocabulary. We finally present a comparison with the clustering induced when estimating the parameters by Gibbs Sampling.

This work has been supported by France Télécom, Division R&D, under contract n°42541441.

2. THE MODEL

We denote by n_D, n_W and n_T, respectively, the number of documents, the size of the vocabulary and the number of themes (that is, the number of components of the mixture model). Since we use a bag-of-words representation of each document, the corpus is fully determined by the count matrix $C = (C_d(w))_{d=1\ldots n_D, w=1\ldots n_W}$, where the notation C_d is used to refer to the word count vector of a specific document d. The multinomial mixture model is such that:

$$\mathrm{P}(C_d; \alpha, \beta) = \sum_{t=1}^{n_T} \alpha_t \frac{l_d!}{\prod_{w=1}^{n_W} C_d(w)!} \prod_{w=1}^{n_W} \beta_{wt}^{C_d(w)} \qquad (1)$$

which corresponds to the following probabilistic generative mechanism:

1. sample a theme t in $\{1, \ldots, n_T\}$ with probabilities $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_{n_T})$
2. sample l_d (which denotes the length of document d) words from a multinomial distribution with parameter $(l_d;\ \beta_{1t}, \beta_{2t}, \ldots, \beta_{n_W t})$

The notation β is used to denote the collection of theme-specific word probabilities. Note that the document length itself is taken as an exogenous variable and its distribution is not accounted for in the model. As all documents are assumed to be independent, the corpus log-likelihood $\mathcal{L}$ is given by $\sum_{d=1}^{n_D} \log \mathrm{P}(C_d; \alpha, \beta)$.

To estimate the model parameters, we use the Expectation-Maximization (EM) algorithm with independent non-informative Dirichlet priors on α (with hyperparameter θ_α) and on the columns $\beta_{\bullet t}$, for $t = 1, \ldots, n_T$, (with hyperparameter θ_β). Denoting the current estimates of the parameters by α' and β' and the latent (unobservable) theme of document d by T_d, it is straightforward to check that each iteration of the EM algorithm updates the parameters ac-

cording to:

$$\mathrm{P}(T_d = t|C;\alpha',\beta') = \frac{\alpha'_t \prod_{w=1}^{n_W} \beta'^{C_d(w)}_{wt}}{\sum_{t'=1}^{n_T} \alpha'_{t'} \prod_{w=1}^{n_W} \beta'^{C_d(w)}_{wt'}} \quad (2)$$

$$\alpha_t \propto \theta_\alpha - 1 + \sum_{d=1}^{n_D} \mathrm{P}(T_d = t|C;\alpha',\beta') \quad (3)$$

$$\beta_{wt} \propto \theta_\beta - 1 + \sum_{d=1}^{n_D} C_d(w)\,\mathrm{P}(T_d = t|C;\alpha',\beta') \quad (4)$$

where the normalization factors are determined by the constraints $\sum_{t=1}^{n_T} \alpha_t = 1$ and $\sum_{w=1}^{n_W} \beta_{wt} = 1$, for t in $\{1,\ldots,n_T\}$. In experiments presented in [5], we observed that changing the values of $\theta_\alpha - 1$ and $\theta_\beta - 1$ did not make the most important differences in the results. Thus, in the rest of this article, we set them respectively to 0 and 0.1.

3. EXPERIMENTAL FRAMEWORK

We selected 1,600 texts[1] from the 2000 Reuters Corpus, from four well-defined categories (sports, health, disasters, employment). All experiments are performed using four-fold cross-validation (with 4 random splits of the corpus). As will be seen below, initialization of the EM algorithm does play a very important role in obtaining meaningful document clusters. To evaluate the performance of the model, one option is to look at the value of the log-likelihood at the end of the learning phase. However, this measure is available only on the training data and does not tell us anything about the generalization abilities of the model. In the context of language processing, its counterpart on the test data is generally expressed in terms of *perplexity*:

$$\widehat{\mathcal{P}}^\star = \exp[-\frac{1}{l^\star}\sum_{d=1}^{n_D^\star} \log(\sum_{t=1}^{n_T} \alpha_t \prod_{w=1}^{n_W} \beta_{wt}^{C_d^\star(w)})]\,,$$

which quantifies how much the model is able to predict new data, generically denoted by the star superscript. The normalization by the total number of word occurrences $l^\star$ in the test corpus $C^\star$ is conventional and used to allow comparison with simpler models such as the unigram model, which ignores the document level. A second indicator, also computable on the test data, is the *mutual information* between the clustering produced by the model and the Reuters categories, which is more directly related to our ability to accurately cluster the data, or at least to recover the original clustering. It is defined as:

$$\widehat{\mathcal{MI}}^\star = \sum_{c=1}^{n_C}\sum_{t=1}^{n_T}(\frac{1}{n_D^\star}\sum_{d=1}^{n_D^\star} \mathrm{P}(\Gamma_c = c|C_d^\star)\,\mathrm{P}(T_d = t|C_d^\star))$$
$$\times \log \frac{n_D^\star \sum_{d=1}^{n_D^\star} \mathrm{P}(\Gamma_c = c|C_d^\star) p(T_d = t|C_d^\star)}{(\sum_{d=1}^{n_D^\star} \mathrm{P}(\Gamma_c = c|C_d^\star))(\sum_{d=1}^{n_D^\star} \mathrm{P}(T_d = t|C_d^\star))}$$

where $\mathrm{P}(\Gamma_c = c|C_d)$ is the "probability" that document d belongs to category c (usually 0 or 1, as most documents belong to a unique Reuters category) and $\mathrm{P}(T_d = t|C_d)$ is the output of the model (probability that the document d belongs to theme t). The estimated mutual information is then normalized, respectively, by the marginal entropies of the themes and categories. The harmonic average of those scores (between 0 and 1) is referred to as the *(MI) F-Score*.

4. IMPORTANCE OF INITIALIZATION

After a bit of experimentation, we found that a good option is to make sure that, initially, all clusters overlap significantly and that none of the theme-dependent word probabilities is too small. The Dirichlet initialization thus consists in sampling an initial (fictitious) configuration of posterior probabilities in (2) which is close to equiprobability[2].

To get an idea about the best achievable performance, we also used the Reuters categories as initialization. We establish a one-to-one mapping between the mixture components and the Reuters categories, setting for every document the initial posterior probability in (2) to 1 for a given theme. Figure 1 displays the corresponding training data likelihood and test data perplexity as functions of the number of iterations. As the results are averaged over several folds and runs, we plot here the mean over the different experiments and the values of the mean plus or minus the standard deviation (respectively with downward and upward triangles). The first striking observation is that the gap between both initializations is huge. With the Dirichlet initialization, we are able to predict the word distribution more accurately than with the unigram model but much worse than with the somewhat ideal initialization. This gap is also patent for the training data log-likelihood and the Mutual Information F-Score, not shown here, but with a final value of 0.87 for the Reuters initialization and an average around 0.25 for the Dirichlet initialization. To get an idea of the signification of these numbers, we randomly perturbated a certain amount of the Reuters tags and computed the MI F-Score with the original categorization. Proceeding this way, perturbing (respectively) 5%, 15% and 50% of the document labels gives

[1] This relatively small size is dictated by the need to conduct a large number of iterations to get meaningful results with Gibbs sampling.

[2] It is not possible to start with exact equiprobability, or, else, it can be seen from the update equations that all word distributions remain similar and the clusters never separate from one another. Hence, we sample from a Dirichlet distribution with the same parameter for all words, so that word probabilities are *a priori* distributed from an exchangeable distribution.

F-Score of 0.9, 0.7 and 0.25. Hence 0.25 corresponds to a rather poor performance.

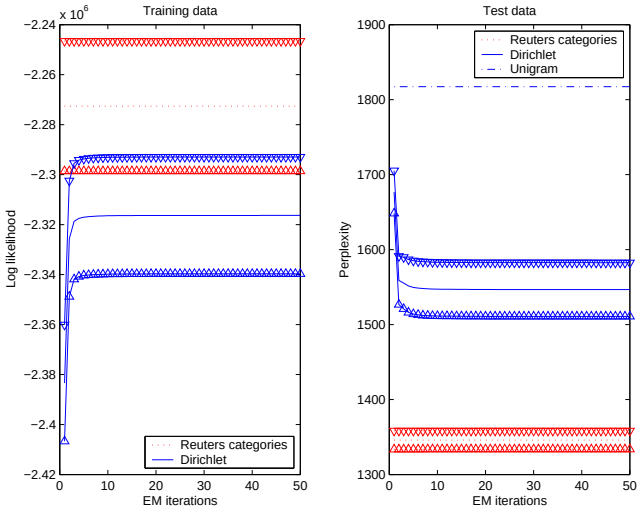

Fig. 1. Evolution of Log-likelihood and Perplexity over the EM iterations.

As the Dirichlet initialization involves random sampling, it may be of interest to check how the performance changes from one run to another. We report the values of log-likelihood and MI F-Score for various runs, without averaging them, in Figure 2. Although the differences from one run to another are striking in terms either of log-likelihood or of quality of the clustering produced, we are always much below the performance obtained with the Reuters categories initialization. In the rest of this article, we represent log-likelihood on the training data and mutual information F-Score on the test data for different experiments. We do not represent perplexity curves, even though they are commonly used in textual data analysis, since the MI F-Score evaluates more directly the model performance for the task we are ultimately interested in. Depending on the readability of the results, we either plot all runs, as in Figure 2, or their average and standard deviation, as in Figure 1.

5. HIERARCHICAL CLUSTERING

Faced to a clustering problem where the final number of components is unknown, it is common to try first to find the most meaningful few clusters and then iteratively perturb the groups obtained to split them into several new clusters. This “divide-and-conquer” approach intends to treat the problem step by step in order to make the whole process easier.

In our case, one way to do that is to use the possibility to initialize the algorithm on the posterior probabilities of a document to belong to a theme. In our case, there are only

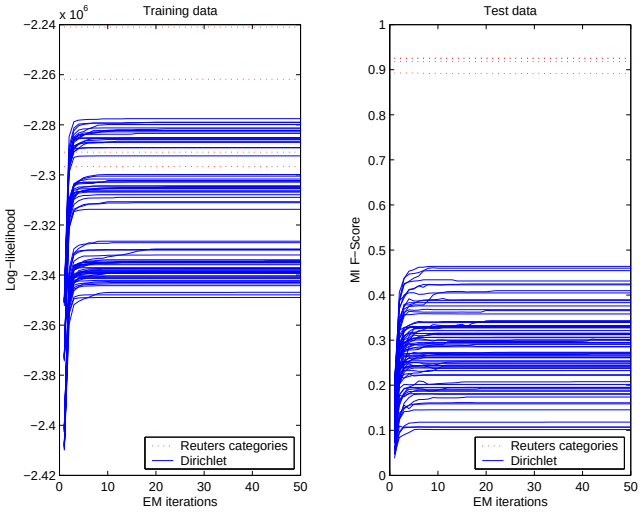

Fig. 2. Evolution of Log-likelihood and MI F-Score over the EM iterations for different Dirichlet initializations.

two rounds of iterations since the final number of themes is 4. We start with two themes, with the initialization Dirichlet and get two posterior probabilities distributions on the documents, one for each theme. In the next rounds, 4 themes are used, theme 1 and 2 being initialized from a Dirichlet sampling from the distribution of the first theme of the previous round and theme 3 and 4 deriving similarly from the second theme of the previous round. Since they involve a random initialization, the experiments are repeated 10 times on each fold. We plot all runs in Figure 3. The graph clearly shows when the split is performed (after the 25th iteration of the EM algorithm).

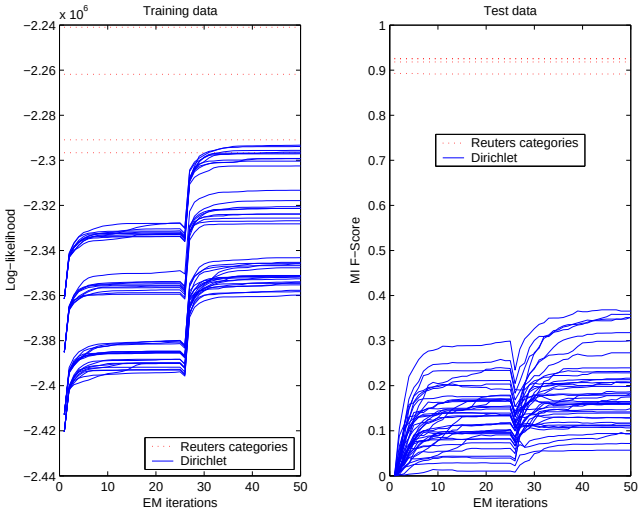

Fig. 3. Evolution of Log-likelihood and MI F-Score over the EM iterations with hierarchical clustering.

The results are disappointing but this experiment is above all interesting because of the disagreement between the training data log-likelihood and test data MI F-Score. Hierarchical clustering does perform similarly to Dirichlet initialization in terms of log-likelihood (leftmost graph in Figure 3 to be compared with corresponding plot in Figure 2). However, when looking at the Mutual Information on the test data, most results are below 0.3, which indeed corresponds to a very unsatisfactory clustering as discussed above. The performance is thus worse than with the basic approach.

The basic idea behind the use of hierarchical clustering is that reducing the number of parameters should improve the quality of the inference. However, reducing the number of themes does not make the matrix β significantly smaller in our case, as the word dimension is, by far, the largest one. In addition the hierarchical approach may have the drawback that it forces unnatural groupings in the initial phase when only two clusters are used. We now turn to another (more drastic) way to reduce the size of the problem by looking at the other dimension of β: the number of words in vocabulary.

6. INFLUENCE OF THE VOCABULARY SIZE

In the experiments conducted to assess the influence of the smoothing parameter θ_β [5], we observed that more smoothing slightly improved the results with the Dirichlet initialization but not with the Reuters categories initialization. We analyzed this fact as a hint that the rarest words were helpful only when properly initialized[3]. Hence, an interesting experiment is to check how the algorithm behaves with the Dirichlet initialization and with only the most frequent words.

We now adjust the vocabulary size by removing rare words. The results in Figure 4 suggest that, on the best runs, we can substantially improve the performance of the model with the Dirichlet initialization. We believe this is an effect of the so called "curse-of-dimensionality" phenomenon: with full vocabulary, the size of the vector space (≈25,000 words) seems too large with respect to the number of training documents (1,200). Figure 5 shows that the somewhat optimal number of words, as far as the (MI) F-Score is concerned, seems to be precisely 500.

7. A HEURISTIC INFERENCE METHOD

The experiments previously described brought up two important points:

- Log-likelihood at the end of the training phase is a reasonable indicator of the quality of the clustering in terms of mutual information. Now, unlike the log-likelihood, which we are able to compute as soon as we have the count matrix and an estimate of the parameters, the MI F-Score is not accessible in a real-world problem since we obviously ignore what is the best clustering. Therefore, it is particularly interesting to have an approximate correlation between a measure available at training phase and the final result we are ultimately interested in.

- Learning parameters on smaller vocabularies yields better results, in the sense of reducing the gap with the ideal initialization, than using all the words from the start. After several EM iterations, we thus have values for posterior probabilities of a text to belong to a given theme, from equation 2, and we know from Section 6 that they induce a good clustering on the corpus.

The main idea of our heuristic inference method is to obtain "good" posterior probabilities with a small vocabulary and to use them as initialization for a new round of EM

[3]In effect, increasing the smoothing parameter leads to homogenize the way we deal with rare words, regardless of the number of times they occur in the training data.

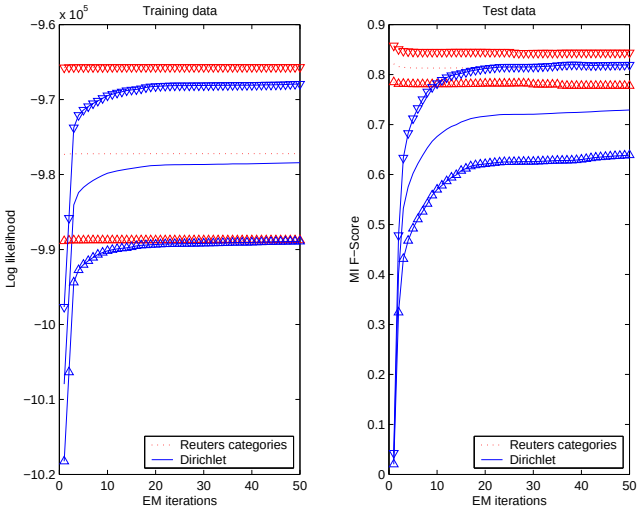

Fig. 4. Evolution of Log-likelihood and MI F-Score over the EM iterations with a vocabulary of size 500.

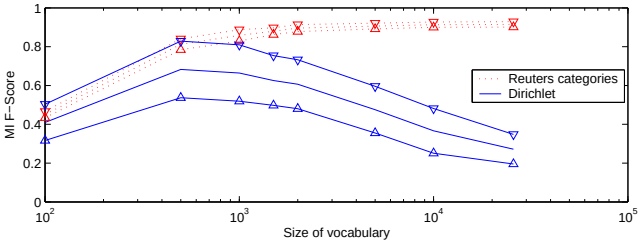

Fig. 5. Mutual Information after the last EM iteration, as a function of the vocabulary size.

iterations, with a larger vocabulary. Thus we avoid the problem of not knowing how to initialize the β parameters corresponding to rare words since we start from the other step of the algorithm (the "M" step). When the vocabulary size is increased, the probabilities associated with new words are thus implicitly initialized on their average count in the corpus, weighted by the current posterior probabilities.

To sum up, the pseudo-code for the algorithm is:

```
vocabsizes = [500 1500 5000 10000 25775]
postprob = group of nD x nT stochastic
           matrices initialized on the
           constant matrix of general
           term 1/nT
for i = 1 to length(vocabsizes)
  vocabulary = most frequent
               vocabsizes(i) words
  for j = 1 to number of runs
    initprob = sample Dirichlet variables
               centered on the distri-
               butions in postprob
    run iterations of the EM algorithm
      starting from initprob
    save final posterior probabilities
      and corresponding log-likelihood
  end
  postprob = keep new posterior proba-
             bilities yielding the best
             likelihoods
end
```

The initial size of vocabulary (500) was chosen according to the results reported in Figure 5. The other sizes were set to get an increase approximately regular in the total number of occurrences in the count matrix from one step to the next. We do not describe here the other parameters (size of the group postprob, number of runs and so on) and the details of the algorithm (such as the sampling from one of the distributions in postprob).

The results are shown in Figure 6 in terms of mean and standard deviation[4] at the end of the last EM iteration. We represent the different steps (or equivalently the different sizes of vocabulary since we add new terms at every step) along the horizontal axis. We note a major improvement in terms of log-likelihood, managing to outperform slightly the reference Reuters Categories initialization (in the figure, the curves are almost superposed), reaching a level far above the initial random Dirichlet sampling on the full vocabulary in Figure 2. If we now turn to the extrinsic application which consists in recovering the (human made) clustering that comes with Reuters database, as measured by the MI F-Score, the algorithm does not outperform the ideal initialization but its performance is much better than before, the mean being around 0.85, to compare with the average of Figure 2, around 0.25. From these experiments, the benefit of learning the values of the parameters corresponding to different parts of the vocabulary incrementally is clear.

[4]As in the previous experiments, to make up for the effects of random initializations, several runs were conducted for each fold. However, the results with this method seem much more stable than with the initial Dirichlet initialization.

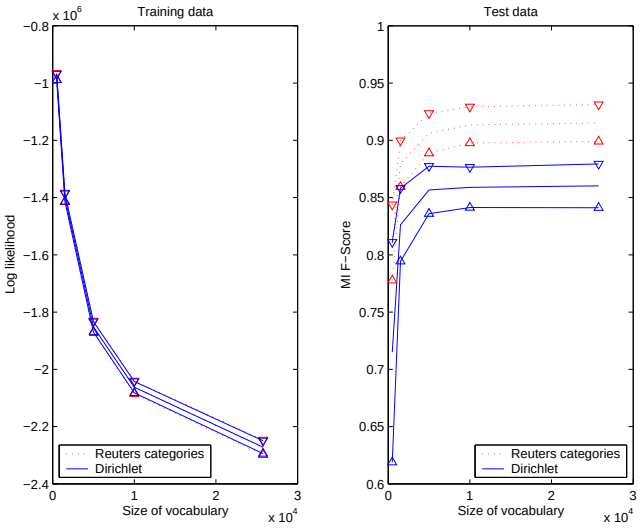

Fig. 6. Evolution of Log-likelihood and MI F-Score over the different steps of a heuristic algorithm.

8. GIBBS SAMPLING

In this last section, we experiment with an MCMC inference method, Gibbs Sampling, which has been successfully applied to LDA, for instance in [6]. We repeatedly:

- sample a theme indicator in $\{1, \ldots, n_T\}$ for each document from a multinomial distribution whose parameter is given by the posterior probability that the document belongs to each of the themes;
- sample values for α, β which, conditionally upon the theme indicators, follow Dirichlet distributions;
- compute new posterior probabilities according to (2).

Unfortunately, the number of iterations typically needed to guarantee a good exploration of the space is much larger than with the EM algorithm. Therefore, we only ran the algorithm five times on fold 1. In Figure 7 we report only the first 20,000 iterations as running the algorithm longer brought no substantial improvement. The performance is varying a lot from one run to another and, occasionally, large changes occur during a particular run with striking consequences on both the log-likelihood and the MI F-Score. This behavior clearly shows that, in this context, the Gibbs sampler does not really attain its objective as it gets trapped, as the EM algorithm, in local modes. Hence, one does

not really simulate from the actual posterior distribution but rather from the posterior restricted to a rather "small" subset of the space of latent variables and parameters. It is interesting to note that, while the results in terms of log-likelihood are in the same range as with the EM algorithm (Figure 2), the values obtained for mutual information are much better. Regarding all measures, the performance is anyway several levels below the one obtained with the ad-hoc inference method of Section 7.

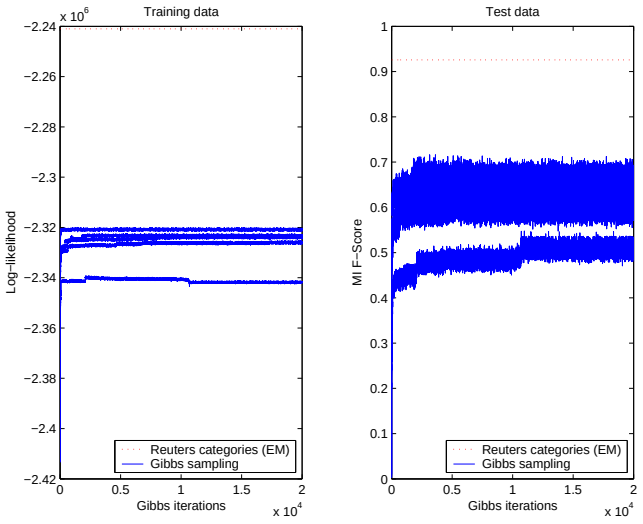

Fig. 7. Evolution of Log-likelihood and MI F-Score over the Gibbs Sampling iterations with full vocabulary.

9. CONCLUSION

In this article, we have presented several methods for the inference of the parameters of a simple mixture of multinomial models for text mining. An evaluation framework based on several measures allowed us to understand the discrepancy between the performance typically obtained with a single run of the EM algorithm and the best scores we could possibly attain when initializing on a somewhat ideal clustering.

We tried various methods to reduce this gap, from hierarchical clustering to Gibbs sampling and other heuristic ideas based on the specificity of the problem, such as learning different parts of the vocabulary incrementally. Reducing the number of words as well as inferring with Gibbs sampling lead to an improvement in comparison to the most straightforward application of the EM algorithm. The best option consists in running several different trials with various initializations, keeping the best ones according to their log-likelihood and incrementally increasing the size of vocabulary.

10. REFERENCES

[1] Thomas Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning Journal*, vol. 42, no. 1, pp. 177–196, 2001.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet allocation," in *Advances in Neural Information Processing Systems (NIPS)*, Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, Eds., Cambridge, MA, 2002, vol. 14, pp. 601–608, MIT Press.

[3] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom M. Mitchell, "Text classification from labeled and unlabeled documents using EM," *Machine Learning*, vol. 39, no. 2/3, pp. 103–134, 2000.

[4] Fabrice Clérot, Olivier Collin, Olivier Cappé, and Eric Moulines, "Le modèle "monomaniaque" : un modèle statistique simple pour l'analyse exploratoire d'un corpus de textes," in *Colloque International sur la Fouille de Texte (CIFT'04)*, La Rochelle, 2004.

[5] Loïs Rigouste, Olivier Cappé, and François Yvon, "Evaluation of a probabilistic method for unsupervised text clustering," in *International Symposium on Applied Stochastic Models and Data Analysis (ASMDA)*, 2005.

[6] Thomas L. Griffiths and Mark Steyvers, "A probabilistic approach to semantic representation," in *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.