Available online at www.sciencedirect.com
ScienceDirect
Information Processing and Management 43 (2007) 1260–1280
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Inference and evaluation of the multinomial mixture model for text clustering

Loïs Rigouste, Olivier Cappé, François Yvon *

GET/Télécom Paris & CNRS/LTCI, 46 rue Barrault, 75634 Paris Cedex 13, France

Received 26 June 2006; received in revised form 30 October 2006; accepted 4 November 2006
Available online 4 January 2007

Abstract

In this article, we investigate the use of a probabilistic model for unsupervised clustering in text collections. Unsupervised clustering has become a basic module for many intelligent text processing applications, such as information retrieval, text classification or information extraction.

Recent proposals have been made of probabilistic clustering models, which build "soft" theme-document associations. These models allow to compute, for each document, a probability vector whose values can be interpreted as the strength of the association between documents and clusters. As such, these vectors can also serve to project texts into a lower-dimensional "semantic" space. These models however pose non-trivial estimation problems, which are aggravated by the very high dimensionality of the parameter space.

The model considered in this paper consists of a mixture of multinomial distributions over the word counts, each component corresponding to a different theme. We propose a systematic evaluation framework to contrast various estimation procedures for this model. Starting with the expectation-maximization (EM) algorithm as the basic tool for inference, we discuss the importance of initialization and the influence of other features, such as the smoothing strategy or the size of the vocabulary, thereby illustrating the difficulties incurred by the high dimensionality of the parameter space. We empirically show that, in the case of text processing, these difficulties can be alleviated by introducing the vocabulary incrementally, due to the specific profile of the word count distributions. Using the fact that the model parameters can be analytically integrated out, we finally show that Gibbs sampling on the theme configurations is tractable and compares favorably to the basic EM approach.

Keywords: Multinomial mixture model; Expectation-maximization; Gibbs sampling; Text clustering

1. Introduction

The wide availability of huge collections of text documents (news corpora, e-mails, web pages, scientific articles...) has fostered the need for efficient text mining tools. Information retrieval, text filtering and

* Corresponding author. Tel.: +33 1 45 81 77 59; fax: +33 1 45 81 31 19.
E-mail addresses: rigouste@enst.fr (L. Rigouste), cappe@enst.fr (O. Cappé), yvon@enst.fr (F. Yvon).

0306-4573/$ - see front matter
doi:10.1016/j.ipm.2006.11.001

classification, and information extraction technologies are rapidly becoming key components of modern information processing systems, helping end-users to select, visualize and shape their informational environment.

Information retrieval technologies seek to rank documents according to their relevance with respect to users queries, or more generally to users informational needs. Filtering and routing technologies have the potential to automatically dispatch documents to the appropriate reader, to arrange incoming documents in the proper folder or directory, possibly rejecting undesirable entries. Information extraction technologies, including automatic summarization techniques, have the additional potential to reduce the burden of a full reading of texts or messages. Most of these applications take advantage of (unsupervised) *clustering techniques* of documents or of document fragments: the unsupervised structuring of documents collections can for instance facilitate its indexing or search; clustering a set of documents in response to a user query can greatly ease its visualization; considering sub-classes induced in a non-supervised fashion can also improve text classification (Vinot & Yvon, 2003), etc. Tools for building thematically coherent sets of documents are thus emerging as a basic technological block of an increasing number of text processing applications.

Text clustering tools are easily conceived if one adopts, as is commonly done, a *bag-of-word* representation of documents: under this view, each text is represented as a high-dimensional vector which merely stores the counts of each word in the document, or a transform thereof. Once documents are turned into such kind of numerical representation, a large number of clustering techniques become available (Jain, Murphy, & Flynn, 1999) which allow to group documents based on ''semantic'' or ''thematic'' similarity. For text clustering tasks, a number of proposal have recently been made which aim at identifying probabilistic (''soft'') theme-document associations (see, e.g., Blei, Ng, & Jordan, 2002; Buntine & Jakulin, 2004; Hofmann, 2001). These probabilistic clustering techniques compute, for each document, a probability vector whose values can be interpreted as the strength of the association between documents and clusters. As such, these vectors can also serve to project texts into a lower-dimensional space, whose dimension is the number of clusters. These probabilistic approaches are certainly appealing, as the projections they build have a clear, probabilistic interpretation; this is in sharp contrast with alternative projection techniques for text documents, such as latent semantic analysis (LSA) (Deerwester, Dumais, Landauer, Furnas, & Harshman, 1990) or non-negative matrix factorization (NMF) techniques (Shahnaz, Berry, Pauca, & Plemmons, 2006; Xu, Liu, & Gong, 2003).

In this paper, we focus on a simpler probabilistic model, in which the corpus is represented by a mixture of multinomial distributions, each component corresponding to a different ''theme'' (Nigam, McCallum, Thrun, & Mitchell, 2000). This model is the unsupervised counterpart of the popular ''Naive Bayes'' model for text classification (see, e.g., Lewis, 1998; McCallum & Nigam, 1998). Our main objective is to analyze the estimation procedures that can be used to infer the model parameters, and to understand precisely the behavior of these estimation procedures when faced with high-dimensional parameter spaces. This situation is typical of the bag-of-word model of text documents but may certainly occur in other contexts (bioinformatics, image processing...). Our contribution is thus twofold:

- We present a comprehensive review of the model and of the estimation procedures that are associated with this model, and introduce novel variants thereof, which seem to yield better estimates for high-dimensional models, and report a detailed experimental analysis of their performance.
- These analyses are supported by a methodological contribution on the delicate, and often overlooked, issue of performance evaluation of clustering algorithms (see, e.g., Halkidi, Batistakis, & Vazirgiannis, 2001). Our proposal here is to focus on a ''pure'' clustering tasks, where the number of themes (the number of dimensions in the ''semantic'' space) is limited, which allows in our case a direct comparison with a reference (manual) clustering.

This article is organized as follows. We firstly introduce the model and notations used throughout the paper. Dirichlet priors are set on the parameters and we may use the expectation-maximization (EM) algorithm to obtain maximum a posteriori (MAP) estimates of the parameters. An alternative inference strategy uses simulation techniques (Markov Chain Monte Carlo) and consists in identifying conditional distributions from which to generate samples. We show, in Section 2.3, that it is possible to marginalize analytically all continuous parameters (thematic probabilities and theme-specific word probabilities). This result generalizes an observation that was used, in the context of the latent Dirichlet allocation (LDA) model by Griffiths and