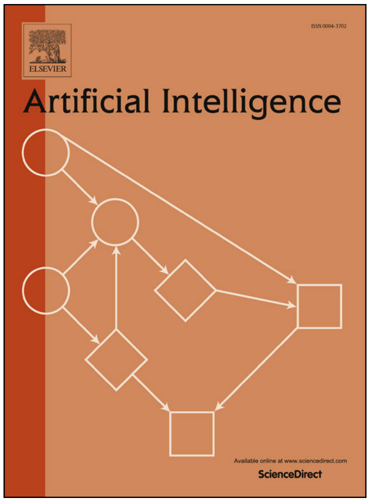

Latent tree models for hierarchical topic detection

Peixian Chen, Nevin L. Zhang, Tengfei Liu, Leonard K.M. Poon, Zhourong Chen,
Farhan Khawar

PII: S0004-3702(17)30073-5
DOI: http://dx.doi.org/10.1016/j.artint.2017.06.004
Reference: ARTINT 3020

To appear in: *Artificial Intelligence*

Received date: 2 May 2016
Revised date: 18 June 2017
Accepted date: 26 June 2017

Please cite this article in press as: P. Chen et al., Latent tree models for hierarchical topic detection, *Artif. Intell.* (2017), http://dx.doi.org/10.1016/j.artint.2017.06.004

Latent Tree Models for Hierarchical Topic Detection

Peixian Chen[a], Nevin L. Zhang[a,*], Tengfei Liu[b], Leonard K. M. Poon[c], Zhourong Chen[a], Farhan Khawar[a]

[a] *Department of Computer Science and Engineering*
The Hong Kong University of Science and Technology, Hong Kong
[b] *Ant Financial Services Group, Shanghai*
[c] *Department of Mathematics and Information Technology*
The Education University of Hong Kong, Hong Kong

Abstract

We present a novel method for hierarchical topic detection where topics are obtained by clustering documents in multiple ways. Specifically, we model document collections using a class of graphical models called *hierarchical latent tree models (HLTMs)*. The variables at the bottom level of an HLTM are observed binary variables that represent the presence/absence of words in a document. The variables at other levels are binary latent variables that represent word co-occurrence patterns or co-occurrences of such patterns. Each latent variable gives a soft partition of the documents, and document clusters in the partitions are interpreted as topics. Latent variables at high levels of the hierarchy capture long-range word co-occurrence patterns and hence give thematically more general topics, while those at low levels of the hierarchy capture short-range word co-occurrence patterns and give thematically more specific topics. In comparison with LDA-based methods, a key advantage of the new method is that it represents co-occurrence patterns explicitly using model structures. Extensive empirical results show that the new method significantly outperforms the LDA-based methods in term of model quality and meaningfulness of topics and topic hierarchies.

[*] Corresponding author
Email address: `lzhang@cse.ust.hk` (Nevin L. Zhang)