Topic Browsing System for Research Papers Based on Hierarchical Latent Tree Analysis

Leonard K. M. Poon[1], Chun Fai Leung[2], Peixian Chen[2], and Nevin L. Zhang[2]

[1] Department of Mathematics and Information Technology
The Education University of Hong Kong, Hong Kong SAR, China
kmpoon@eduhk.hk
[2] Department of Computer Science and Engineering
The Hong Kong University of Science and Technology, Hong Kong SAR, China
cfleungac@connect.ust.hk, pchenac@cse.ust.hk, lzhang@cse.ust.hk

Abstract. New academic papers appear rapidly in the literature nowadays. This poses a challenge for researchers who are trying to keep up with a given field, especially those who are new to a field and may not know where to start from. To address this kind of problems, we have developed a topic browsing system for research papers where the papers have been automatically categorized by a probabilistic topic model. Rather than using Latent Dirichlet Allocation (LDA) for topic modeling, we use a recently proposed method called hierarchical latent tree analysis, which has been shown to perform better than some state-of-the-art LDA-based methods. The resulting topic model contains a hierarchy of topics so that users can browse topics at different levels. The topic model contains a manageable number of general topics at the top level and allows thousands of fine-grained topics at the bottom level.

1 Introduction

New academic papers appear rapidly in the literature nowadays. This makes a good contribution to the acquisition of knowledge but poses a challenge for researchers trying to keep up with a given field. The problem may be even worse for postgraduate students who are new to a field and may not know where to start from. Researchers usually use keywords to search for related papers using a search engine. They may then group related papers together to find out the main topics in the field. This process can be time-consuming.

The above approach can be regarded as a bottom-up approach. In contrast, a top-down approach would start with topic hierarchy. Researchers can then pick a general topic and drill down to find more specific topics. Papers related to any of the topics can be presented to the researchers when requested.

To allow the top-down approach, traditionally a taxonomy has to be defined manually. Papers are then be categorized manually according to the taxonomy. There are two issues with the traditional method. First, it requires much effort. Second, the topics in the taxonomy may not be able to keep up with latest development. To address those issues, we propose a topic browsing system using a recent method for automatically building a topic tree and categorizing papers.

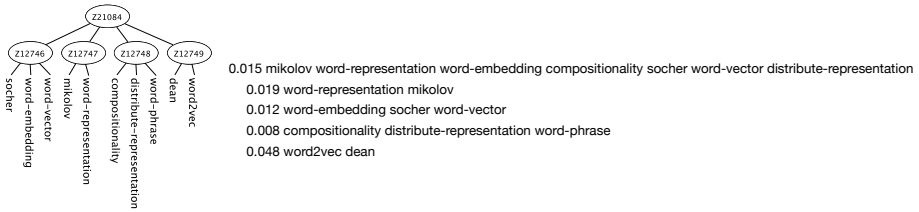

Fig. 1: A latent tree model (left) and the topic hierarchy extracted from it (right).

2 Hierarchical Latent Tree Analysis

Topic models are often used to categorize documents automatically. They can detect topics from a collection of documents and classify each document according to the detected topics. We have recently developed a topic modeling method called *hierarchical latent tree analysis* (HLTA) [3]. Unlike Latent Dirichlet Allocation [2], HLTA produces a hierarchy of topics. Our recent work [3] has also shown that HLTA produces topics and topic hierarchies of better quality than two state-of-the-art LDA-based methods, namely hLDA [1] and nHDP [5].

HLTA uses a class of tree-structured probabilistic graphical models called *latent tree models* (LTMs) [4]. Fig 1 shows an example of LTM for topic modeling. In the model, the leaf nodes (unframed nodes) represent observed word variables $\boldsymbol{W}$, whereas its internal nodes (oval nodes) represent unobserved topic variables $\boldsymbol{Z}$. All variables are binary. Each word variable $W \in \boldsymbol{W}$ indicates the presence or absence of the corresponding word in a document. Each topic variable $Z \in \boldsymbol{Z}$ indicates whether a document belongs to the corresponding topic.

An LTM can be regarded as a Bayesian network by rooting at one of its latent nodes. Let $pa(X)$ be the parent of a variable X.Then the LTM defines a joint distribution over all observed and latent variables as follows: $P(\boldsymbol{W}, \boldsymbol{Z}) = \prod_{X \in \boldsymbol{W} \cup \boldsymbol{Z}} P(X|pa(X))$.

Denote a document by $d = (w_1, \ldots, w_M)$, where w_i is the observed value of word variable $W_i \in \boldsymbol{W}$. Whether a document d belongs to a topic $Z \in \boldsymbol{Z}$ can be determined by the probability $P(Z|d)$. The LTM gives a *multi-membership model* since a document can belong to multiple topics. And unlike in LDA, the topic probabilities $P(Z|d)$ in LTM do not necessarily sum to one.

3 System Overview

We have developed a topic browsing system for research papers where the papers can be automatically categorized by a topic model built with HLTA. At the time of writing, the system has categorized 24,307 papers in the field of artificial intelligence published in 7 major conferences and 3 journals from 2000 to 2017. The resulting topic model contains a hierarchy of topics. The top level contains a manageable number of general topics and the bottom level contains thousands of fine-grained topics.

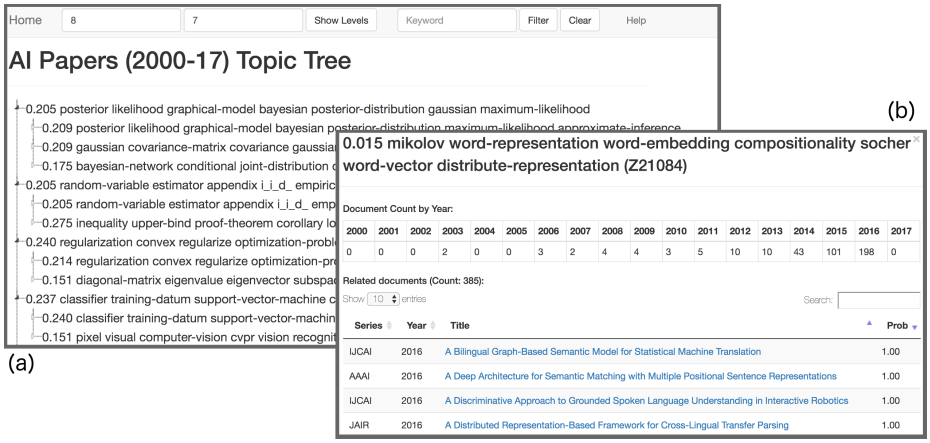

Fig. 2: (a) Home of topic browsing system. (b) Papers for a selected topic.

The home page of our topic browsing system displays a hierarchy of topics (Fig 2a). When a topic node is clicked, a list of papers belonging to that topic is displayed (Fig 2b). The system is built as a Node.js web application. Papers information is stored in a MongoDB database and is loaded through Ajax.

To prepare the data, papers are downloaded from related proceedings and journal websites. The text is preprocessed with word lemmatization, word normalization, and removal of stop words. Then, text is converted to data with bag-of-words representation. We consider n-grams, where $1 \leq n \leq 4$. We use the 10,000 n-grams with highest TF-IDF and appearing in less than 25% of papers.

The converted data is used as input to HLTA, resulting in an LTM. The number of levels and the number of latent variables in each level are automatically determined. A topic hierarchy is extracted based on the tree structure of the LTM, with each internal node representing a topic. Each topic is then characterized by those of its descendent words with highest MI with the topic. Fig 1 shows an example of LTM and the hierarchy extracted from it.

The LTM is also used to classify papers according to the topics detected. A paper d is assigned to topic Z if $P(Z = 1|d) > 0.5$. Besides, the size of each topic Z is estimated by the marginal probability $P(Z)$. In the topic hierarchy of Fig 1, the number on each row indicates the topic size.

The code for processing is available at `https://github.com/kmpoon/hlta`.

4 Demonstration

The upper part of the hierarchy in our system is shown on the next page. The topics include: (1) graphical models and Bayesian methods; (2) statistical analysis; (3) optimization; (4) matrix factorization; (5) classification and support vector machines; (6) computer vision; (7) data mining; (8) text mining; (9) agents; (10) reinforcement learning; (11) logic; (12) impact; (13) user interface and temporal

(1) 0.205 posterior likelihood graphical-model bayesian posterior-distribution gaussian maximum-likelihood
(2) 0.205 random-variable estimator appendix i_i_d_ empirical learning-research-submit sample-size
0.240 regularization convex regularize optimization-problem gradient diagonal-matrix eigenvalue
(3) 0.214 regularization convex regularize optimization-problem gradient convex-optimization norm
(4) 0.151 diagonal-matrix eigenvalue eigenvector subspace principal-component-analysis diag high_dimensional
0.237 classifier training-datum support-vector-machine class-label classification-accuracy classification-problem classification-task
(5) 0.240 classifier training-datum support-vector-machine class-label classification-accuracy classify classification-problem
(6) 0.151 pixel visual computer-vision cvpr vision recognition object-recognition
0.188 document information-retrieval text mining datum-mining baseline corpus
(7) 0.161 mining datum-mining knowledge-discovery sigkdd-international-conference-knowledge sigkdd icdm data-mining
(8) 0.177 document text information-retrieval baseline corpus sigir precision-recall
0.151 agent multi_agent initial-state agent-agent multiagent assume-agent optimal-policy
(9) 0.124 agent multi_agent agent-agent multiagent assume-agent multi_agent-system multiagent-system
(10) 0.098 initial-state reward optimal-policy transition reinforcement-learning policy markov-decision-process
(11) 0.168 logic logical propositional semantics negation predicate disjunction
0.227 continue impact earlier effort back initially happen
(12) 0.229 continue impact earlier effort back initially happen
(13) 0.247 interface temporal people dynamic duration cognitive activity
0.052 intelligence-www_elsevier_com_locate_artint front-matter e_mail-address elsevier-rights-reserve revised-form solver satisfiabilit
0.052 intelligence-www_elsevier_com_locate_artint front-matter e_mail-address elsevier-rights-reserve revised-form satisfiability cc
(14) 0.052 intelligence-www_elsevier_com_locate_artint front-matter e_mail-address revised-form elsevier-rights-reserve correspond
(15) 0.117 satisfiability clause assignment satisfiable constraint-satisfaction-problem assignment-variable literal
(16) 0.151 kernel smola kernel-method scho-lkopf vapnik kernel-function statistical-learning-theory
(17) 0.207 solver optimal-solution problem-instance problem-solve search-tree solve-problem search-algorithm

systems; (14) publishing information; (15) satisfiability; (16) kernel methods; and (17) solvers. The upper part includes many major topics and looks reasonable.

To browse the topics, one can go down the hierarchy (Fig 2a) and click a topic node to show a list of related papers (Fig 2b). The papers are sorted in descending order of membership as indicated by $P(Z|d)$. Links to the original papers can be accessed by clicking the titles in the paper list. Numbers of papers for each year are also shown in a table at the top. This shows the trend of the topic.

Our proposed system can be accessed from: `https://ltm.eduhk.hk/papers/`.

Acknowledgment The work was supported by the Education University of Hong Kong under project RG90/2014-2015R and Hong Kong Research Grants Council under grants 16202515 and 16212516.

References

[1] Blei, D.M., Griffiths, T.L., Jordan, M.I.: The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. Journal of the ACM 57(2), 7:1–7:30 (2010)
[2] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
[3] Chen, P., Zhang, N.L., Poon, L.K.M., Chen, Z.: Progressive EM for latent tree models and hierarchical topic detection. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (2016)
[4] Chen, T., Zhang, N.L., Liu, T., Poon, K.M., Wang, Y.: Model-based multidimensional clustering of categorical data. Artificial Intelligence 176, 2246–2269 (2012)
[5] Paisley, J., Wang, C., Blei, D.M., Jordan, M.I.: Nested hierarchical Dirichlet processes. IEEE Transactions on Pattern Analysis and Machine Intelligence 37(2), 256–270 (2015)