Computer Science > Learning

Scalable Text and Link Analysis with Mixed-Topic Link Models

Yaojia Zhu, Xiaoran Yan, Lise Getoor, Cristopher Moore

(Submitted on 28 Mar 2013)

Many data sets contain rich information about objects, as well as pairwise relations between them. For instance, in networks of websites, scientific papers, and other documents, each node has content consisting of a collection of words, as well as hyperlinks or citations to other nodes. In order to perform inference on such data sets, and make predictions and recommendations, it is useful to have models that are able to capture the processes which generate the text at each node and the links between them. In this paper, we combine classic ideas in topic modeling with a variant of the mixed-membership block model recently developed in the statistical physics community. The resulting model has the advantage that its parameters, including the mixture of topics of each document and the resulting overlapping communities, can be inferred with a simple and scalable expectation-maximization algorithm. We test our model on three data sets, performing unsupervised topic classification and link prediction. For both tasks, our model outperforms several existing state-of-the-art methods, achieving higher accuracy with significantly less computation, analyzing a data set with 1.3 million words and 44 thousand links in a few minutes.

Comments:	11 pages, 4 figures
Subjects:	**Learning (cs.LG)**; Information Retrieval (cs.IR); Social and Information Networks (cs.SI); Data Analysis, Statistics and Probability (physics.data-an); Machine Learning (stat.ML)
ACM classes:	G.3; H.3.3; H.4; I.2
Cite as:	**arXiv:1303.7264 [cs.LG]**
	(or **arXiv:1303.7264v1 [cs.LG]** for this version)

Submission history

From: Yaojia Zhu [view email]

[v1] Thu, 28 Mar 2013 22:34:51 GMT (2323kb)

Which authors of this paper are endorsers?