Statistics > Machine Learning

Topic Discovery through Data Dependent and Random Projections

Weicong Ding, Mohammad H. Rohban, Prakash Ishwar, Venkatesh Saligrama

(Submitted on 15 Mar 2013 (v1), last revised 18 Mar 2013 (this version, v2))

We present algorithms for topic modeling based on the geometry of cross-document word-frequency patterns. This perspective gains significance under the so called separability condition. This is a condition on existence of novel-words that are unique to each topic. We present a suite of highly efficient algorithms based on data-dependent and random projections of word-frequency patterns to identify novel words and associated topics. We will also discuss the statistical guarantees of the data-dependent projections method based on two mild assumptions on the prior density of topic document matrix. Our key insight here is that the maximum and minimum values of cross-document frequency patterns projected along any direction are associated with novel words. While our sample complexity bounds for topic recovery are similar to the state-of-art, the computational complexity of our random projection scheme scales linearly with the number of documents and the number of words per document. We present several experiments on synthetic and real-world datasets to demonstrate qualitative and quantitative merits of our scheme.

Subjects: **Machine Learning (stat.ML)**; Learning (cs.LG)
Cite as: **arXiv:1303.3664 [stat.ML]**
(or **arXiv:1303.3664v2 [stat.ML]** for this version)

Submission history

From: Weicong Ding [view email]
[v1] Fri, 15 Mar 2013 02:37:19 GMT (416kb)
[v2] Mon, 18 Mar 2013 13:11:02 GMT (416kb)

Which authors of this paper are endorsers?