氏　　名	THAN QUANG KHOAT			
学位の種類	博士(知識科学)			
学位記番号	博知第 144 号			
学位授与年月日	平成 25 年 9 月 24 日			
論文題目	Fast and Sparse Inference for Topic Models(トピックモデルのための高速スパース推論)			
論文審査委員	主査	Ho Tu Bao	北陸先端科学技術大学院大学	教授
		中森　義輝	同	教授
		溝口　理一郎	同	教授
		Dam Hieu Chi	同	准教授
		鷲尾　隆	大阪大学	教授

論文の内容の要旨

Topic modeling has been increasingly maturing to be an attractive research area. Originally motivated from textual applications, it has been going beyond far from text to touch upon many amazing applications in Computer Vision, Bioinformatics, Software Engineering, Forensics, Cognitive Science, History, Politics, to name a few. It is believed to be one of the keys to automatically understanding documents written by human, and to uncovering how human knowledge is created and represented.

This thesis studies to model texts at a large scale. In other words, the thesis studies to propose models that most appropriately generate documents, and then to derive efficient methods for learning those models from a large number of available texts. To this end, the thesis systematically elucidates the two fundamental issues to be resolved: *inference of topic mixtures* and *model complexity*. The thesis then targets at developing provably fast algorithms that can recover sparse topic mixtures for documents, and developing fast algorithms to learn sparse topic models.

The first contribution is the introduction of a simple framework for inference of sparse topic mixtures, called *FW*, which is general and flexible enough to be employed in admixture models. The framework enjoys the following key theoretical properties: (1) inference provably converges at a linear rate to the optimal solutions; (2) prior knowledge can be easily incorporated into inference; (3) the sparsity level of topic mixtures can be directly controlled; (4) it is easy to trade off sparsity against quality and runtime. Existing inference methods do not own these properties and often work slowly. Those properties are attractive for large scale modeling.

We demonstrate the goodness and flexibility of *FW* by employing it to design novel methods for supervised dimension reduction. When working with very high dimensional problem, it is sometimes beneficial in efficiency and effectiveness to reduce the dimensionality of the problem, but keep or

make better predictiveness of the response variable. The main result of this study is a novel method that can reach state-of-the-art performance while enjoying 30-450 times faster speed than existing methods.

The second contribution is the introduction of *Fully Sparse Topic Model* (FSTM) for modeling large collections of documents. Three key properties of the model are: (i) the inference algorithm converges at a linear rate to the optimal solutions, (ii) it provides a principled way to directly trade off sparsity of solutions against inference quality and running time, (iii) the learning algorithm has low complexity which is near independent of dimensionality. FSTM overcomes many limitations of existing topic models, and has been demonstrated to work qualitatively on real data. The low computational complexity and low model complexity can help us work with large text collections.

The third contribution is the introduction of a fast algorithm for learning Correlated Topic Models (CTM), as well as a theory of *probable convexity* for analyzing convexity of real functions. Previous studies show that posterior inference in nonconjugate models such as CTM is intractable (NP-hard) in the worse case. However, we show that it may not be true in practice. Indeed, by introducing the concept of *probable convexity*, we show that inference of topic mixtures in CTM and many nonconjugate models is tractable in practice. Based on these findings, a novel algorithm is proposed which is surprisingly simple but is easily parallelizable or distributable. By extensive experiments, the algorithm is shown to work significantly faster than existing expensive methods while keeping comparable or better quality of the learned models.

論文審査の結果の要旨

Topic modeling has been increasingly maturing to be an attractive research area. Originally motivated from textual applications, it has been going beyond far from text to touch upon many amazing applications in Computer Vision, Bioinformatics, Software Engineering, Forensics, Cognitive Science, History, Politics, to name a few. It is believed to be one of the keys to automatically understanding documents written by human, and to uncovering how human knowledge is created and represented.

This thesis studies to model data at a large scale. In other words, the thesis studies to propose models that most appropriately generate documents, and then to derive efficient methods for learning those models from a large number of available texts. When facing with very large and complex data, the two fundamental problems to be addressed are inference speed and model complexity. A good solution to these problems would make a significant progress for topic modeling in particular and for Machine Learning in general. Hence, this thesis will attack these two problems in a systematic way.

The first contribution is the introduction of a simple framework for inference in topic models, called FW, which is general and flexible enough to be easily employed in mixture models. The framework enjoys the following key theoretical properties: (1) inference converges at a linear rate to the optimal solutions; (2) prior knowledge can be easily incorporated into inference; (3) the sparsity level of latent representations can be directly controlled; (4) it is easy to trade off sparsity against quality and time. Existing inference methods do not own these properties and often work slowly.

The second contribution is the introduction of Fully Sparse Topic Model (FSTM) for modeling large collections of documents. Three key properties of the model are: (i) the inference algorithm converges at a linear rate to the optimal solutions, (ii) it provides a principled way to directly trade off sparsity of solutions against inference quality and running time, (iii) the learning algorithm has low complexity which is near independent of dimensionality. FSTM overcomes many limitations of existing topic models, and has been demonstrated to work qualitatively on real data.

The third contribution is the introduction of a fast algorithm for learning correlated topic models, as well as a theory that theoretically guarantees the good performance of the algorithm. Modeling the interactions of hidden topics implies that we have to model two levels of unknown factors (i.e. topics and their interactions). Hence derivation of an efficient method for learning requires nontrivial efforts. Our proposed algorithm is surprisingly simple but is easily parallelizable or distributable. By extensive experiments, the algorithm is shown to work significantly faster than existing expensive methods while keeping comparable or better quality of the learned models.

The study has shown the candidate's strong ability of independently conducting the scientific research, which should be sufficient to be a considered for the doctoral degree of knowledge science.