1. Summary

The focus of this paper is about probabilistic topic models. As generative models for documents, they show a basic idea that a document is a mixture of topics and a topic is a probability distribution over words. Thus, documents with different content can be generated by tuning the probability distribution over topics.

Compared to Latent Semantic Analysis (LSA) which is another classic statistical method, they both claim that 1) semantic information can be derived from a word-by-document matrix, and 2) dimensionality reduction is essential for the derivation. However, for *semantic properties* of words and documents, LSA represents them as points in Euclidean space, while the topic model applies the concept of probabilistic topics.

Based on the *bag*-of-words assumption, a topic model deals with generative process and statistical inference problem. With the view that a document is a mixture of topics, probabilistic topic models assumes that a word in a document is generated by first sampling a topic from the *document-topic* distribution (θ) and then choosing a word from the *topic-word* distribution (ϕ). In Section 3, hyperparameter setting (Dirichlet prior α for θ in LDA, Dirichlet prior β for ϕ), interpretations (graphical model with plate notation for *repeated sampling*, geometric interpretation with simplex, matrix factorization interpretation with comparison to LSA) and other applications are reviewed.

To extract topics, Gibbs sampling (a Markov chain Monte Carlo) is easy to implement, efficient, and avoids (?) local convergence problems EM algorithm faces. It *directly* estimates the *posterior* distribution over z (the assignment of word *tokens* to topics) given the observed words w. It shows that words are assigned to topics based on how "likely" the word is for a topic and how "dominant" a topic is in a document. Additionally, θ and ϕ could be approximated using the posterior estimations of vector z (corresponding terms are included in the formula).

In terms of polysemy with topics, this word ambiguity is addressed by observing other less ambiguous words in the *same* context. Note that Gibbs sampling is an *iterative* algorithm where topic assignment of each word token is dependent on that of all others.

For similarity computation, document similarity and word similarity are discussed. Both the *symmetrized* Kullback Leibler (KL) and Jensen-Shannon (JS) divergence functions can work well in real life. In order to retrieve the most relevant documents to a query (in the field of information retrieval applications), different approaches such as 1) accessing similarity between the query and each of the candidate documents, and 2) searching documents that maximizes the conditional probability of the query among candidates. Since *stability* plays an important role (especially for short documents), implementing multiple Gibbs samples with average on similarity function should be reasonable (??). On the other hand, for measuring similarity of words, there are approaches such as inferring relevant conditional *topic* distributions and exploring associative relations among words as if generating responses from one of the words.

2. Comments

As Section 7 says, probabilistic topic models make *explicit* assumptions about the *causal* process responsible for generating a document and use statistical methods to identify the latent structure that underlies a set of words (topics). They share some key assumptions with LSA. But for expressing semantic properties of words and documents, it uses the probabilistic concept of "topic" instead of representing words and documents with points in Euclidean space and

extracting/utilizing *linearly independent c*omponents (“factors”). In this way, these models might be less abstract for people to interpreter. And the probabilistic viewpoint also provides new insight to formalize and understand problems such as similarity computation of documents/words.

3. Questions

(1) In the paper, it mentions that EM algorithm suffers from problems such as local maxima of the likelihood function. I was curious if Gibbs sampling algorithm has similar issue? If not, how could it avoid/resolve it?

(2) I find the discussion on “exchangeability of topics” and “stability of topics” in Section 4 important but not easy to comprehend. For example, it is mentioned that “There is *no* a priori *ordering* on the topics that will make the topics identifiable between or even within runs of the algorithm…therefore, the different samples cannot be averaged at the level of topics”, and that “when topics are used to calculate a statistic which is *invariant* to the ordering of the topics, it becomes possible and even important to average over different Gibbs samples”. What does “order” mean? Why topic from different samples are not constrained to be similar? How does it affect taking average on the Gibbs samples?

(3) To deal with polysemy problem, is it easier for probabilistic topic models than for LSA? It seems that Week 4’s reading paper mentions that the representation of a word/document with a *unique* point results in the distortion (to some extend) for some queries having multiple-meaning words. Since probabilistic topic models use alternative ways (topics) to express word-and-document semantic properties, the cause of polysemy problem seems kind of weakened.