# Review of Topic Modeling Journals

Evan M. Gertis
Msc Computer Science, Georgia Southern University

*Abstract*— **This is a review of journals in the area of using LDA for topic modeling. The journals have discussed the milestones that have been achieved in the area of topic modeling particularly Latent Dirchlet Association. They has also highlighted several applications that can be applied in the area of topic modeling systems. Four articles have been reviewed titled as follows: -**

1) **Reading Tea Leaves: How Humans Interpret Topic Models by Jonathan Chang and others.**
2) **Exploring the Space of Topic Coherence Measures by Michael Rder and others.**
3) **Surveying a suite of algorithms that offer a solution to managing large document archives by David M. Blei**

**The following text describes a brief overview of what these articles have reviewed in this field.**

## I. Reading Tea Leaves: How Humans Interpret Topic Models by Jonathan Chang and others.

### A. Authors

(Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, David M. Blei).

### B. Overview

This paper reviews machine probabilistic models for unsupervised analysis of text. It focuses on providing a predictive model of future text and a latent topic representation of the corpus. The main concern here is that topic models which perform better on held-out likelihood may infer less semantically meaningful topics. The paper presents new quantitative methods for measuring semantic meaning in inferred topics. The authors back the measures with large-scale user studies. These studies show that they capture aspects of the model that are typically undetected by previous measures of model quality based on held-out likelihood.

- Introduction
- Topic models and their evaluations
- Case Study
- Using human judgments to examine the topics
- Experimental results

### C. Components of a Topic Model

*1) Topics:* A topic model is a type of statistical model for discovering the abstract ”topics” that occur in a collection of documents. Topic modeling is frequently used as a text-mining tool for the semantic structures in a text.

*2) Word intrusion:* In a word intrusion task, the subject is presented with six randomly ordered words. The task of the user is to find the word which is out of place. For example, apple as the intruding word in the set (dog, cat, horse, apple, pig, cow).

*3) Centralized Server:* This component maintains the details of all the existing vulnerabilities and current patches. It uses Support Vector Machine algorithms to improvise performance of its web scans.

*4) Topic Intrusion:* Topic intrusion tests whether a topic model's decomposition of documents into a mixture of topics agrees with human judgments of the document's content. The remaining intruder topic is chosen randomly from the other low-probability topics in the model.

### D. Experimental results

Three topic models were studied: probabilistic latent semantic indexing (pSLA), latent Dirichlet allocation (LDA), and the correlated topic model (CTM). pLSI is not fully generative, whereas LDA is. In pLSI $\theta_d$, the topic mixture proportions are a parameter for each document. In LDA $\theta_d$ is treated as a random variaable ddrawn from a Dirichlet prior distribution. In CTM the components of $\theta_d$ are nearly independent. Each model was used on two corpora. The log likelihood/predictive rank was used to determine the best performance. Generally, CTM performed the best, followed by LDA, and pSLI.

### E. Discussion

The results from this paper state that practitioners should focus on developing topic models that focus on evaluations that depend on real-world task performance rather than optimizing likelihood-based measures.

## II. Exploring the Space of Topic Coherence Measures

### A. Authors

(SExploring the Space of Topic Coherence Measures, Andreas Both, Alexander Hinneburg)

### B. Overview

This paper proposes a fraamework that constructs exsisting word based coherence measures aas well aas new ones by combining elementary components. The researchers present combinations of components the ooutperform existing measures with respect to correlation to human ratings. A discussion about how results can be transferred to further applications in the context of text mining and information retrieval is held at the endd.

- explanation of coherence with regards to statements or facts
- understandability andd interpretability of topic models with respect to humans
- description of unifying framework for coherence measures

The researchers start with an explanation of coherence. A statement is said to be coherent, if they support each other. For example, ”the game is a team sport”, ”the game is played with a ball”, ”the game demands great physical efforts”. These statements support each other and are coherent. The researchers propose a unifying framework that spans a configuration space of coherence definitions. Previous researchers ranked words as good, neutral or bad. Then the evaluated topic coherence takes the set of N top words of a topic and sums a confirmation measure over all word pairs.

The researchers in this paper propose a framework for comparing different approaches for evaluating topic coherence. Specifically, an approach spans configuration space of coherence definitions. The results achieved by the researches reveal a coherence measure based on a new combination of approaches that approximates human ratings.

Previous researchers, Newman, used an evaluation method that took the set of top N words, ranked the words on good, bad or neutral. Then a confirmation measure was used to evaluate the topic coherence. This method utilized pointwise mutual informationi (PMI). The largest correlation to human toopic coherrence ratings were found when defining the element of vectors as normalized PMI. Fitelson used a method that calulacted the coherence by using the probability of every single word in the set.

The researchers in this paper propose a framework comprised of four steps. Each word is segmented into a set of pairs of word subsets. Then a set of confirmation measures are determined. Then the word probabilities are computed. Finally, the cross product of the subset of words, confirmation measures and probability is computed to determine the configuration space. Each word is segmented into a subset. The method of probability estimation used by researchers in this paper used Boolean documents to compute word probabilities. Finally, a confirmation measure was used the pairs of words to compute the difference, ratio, and likelihood measures. Indirect computation was used to determine confirmation measures in order to avoid introducing words that have low joint probability. Indirect confirmation was computed as vector similarity. Once the confirmation measures were determined all subset pairs were aggregated to a single coherence score. They showed that their method can cover all existing coherence measures.

Researchers concluded that UCI coherence performs better with NPMI. The best perofrming measure was combined aapproach that used which has been often overlooked. The approach uses the indirect cosine measure with the NPMI (normaalized pointwise mutual information).

Coherence measures can be used to rate the quality of topics computed by topic models. The researchers describe an application of coherence measures used in TopicExplorer. The application of this research would help improve topic quality by hiding topics that are irrelevant to a user.

### *C. Surveying a suite of algorithms that offer a solution to managing large document archives*

This article aims to describe probabilistic topic models. How topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation.

The author describes Latent Dirchlet Allocation through a manual process by highlighting different words relating to data analysis. They describe LDA as a statistical model that reflects the intuition that reflects the intuition that documents exhibit multiple topics.

First a distribution over the topics is chosen, then for each word a topic assignment is chose. Then a word is chosen from the corresponding topic.

The author describes the goal of topic modeling is to automatically discover the topics the topics from a collection of documents.

The example show in the paper took 17,000 articles from Science magazine and identified 100 topics. The application from this example is that topic modeling provides an algorithmic solution to managing, organizing, and annotating large archives of texts.

### *D. LDA and probabilistic models*

In generative probabilistic modeling, data is treated as arising from a generative process that includes hidden variables. This defines what is referred to as a joint probability distribution over booth the observed and hidden random variables. LDA uses $\theta_d$ where $\theta_d$ is the topic in proportion for topic $k$ in the document $d$.

The author uses a probabilistic graphical model for describing LDA. Probabalistic graphical models provide a graphical language for describing families of probability distributions.

LDA was developed to fix a pervious problem with pLSI.

The author describes the posterior computation as the division of the joint distribution by all of the random variables which can be computed for any setting of the hidden variables. The denominator is referred to as the marginal probability of observations. One of the major problems in topic modeling is computing the denominator of the posterior. One of the central research goals is to approximate the denominator.

The Gibbs sampling algorithm is one example of a method for approximating it. Markov chains is defined on the hidden topic variables. Then the algorithm collects samples from the distribution. Then an approximation of distribution is determined.

A comparison between sampling based algorithms and variational methods is made. In variational methods rather than approximate the posterior, a member of the family of distributions from the hidden structure is used —- insert difference here — Both algorithms essentially perform a search over the topic structure.

The author explains how LDA can be used to in population genetics to find ancestral populations in in the genetic ancestry of a sample of individuals. Biologist can use LDA to characterize the patterns in populations as topics.

Topic modeling is described as an emerging field. We can use topic models to organize, summarize, and help users explore large corpora One of the problems that needs to be addressed is comparing topic models. This is referred to as the model checking problem.

One of the future directions for topic modeling is the development of data visualizations tools. Researchers need new ways to explore topics visually. The goal of data visualization in topic modeling is to present data in a way that is easy to interpret.

## III. CONCLUSIONS

The field of artificial intelligence is gaining momentum especially in this new era of advanced computing. various fields such as Information Systems Security are now taking advantage of this field to optimize security in systems and provide more secure networks.

## REFERENCES

[1] Michael Rder, Andreas Both, Alexander Hinneburg, Exploring the Space of Topic Coherence Measures

[2] Jonathan Chang, Jordan Boyd-Graber and Sean Gerrish, Chong Wang, David M. Blei. Reading Tea Leaves: How Humans Interpret Topic Models

[3] David M. BLEI, Probabilistic Topic Models, 2012. Surveying a suite of algorithms that offer a solution to managing large document archives.