**Discovering Canonical Correlations between**

**Topical and Topological Information in Document Networks**

**Abstract**

Document network is a kind of intriguing dataset which can provide both topical (textual content) and topological (relationallink) information. A key point in modeling such datasets is to discover proper denominators beneath the text and link. Most previouswork introduces the assumption that documents closely linked with each other share common latent topics. However, the heterophily(i.e., tendency to link to different others) of nodes is neglected, which is pervasive in social networks. In this paper, we simultaneouslyincorporate community detection and topic modeling in a unified framework, and appeal to Canonical Correlation Analysis (CCA) tocapture the latent semantic correlations between the two heterogeneous factors, *community* and *topic*. Despite of the homophily (i.e.,tendency to link to similar others) or heterophily, CCA can properly capture the inherent correlations which *fit* the dataset itself withoutany prior hypothesis. We also impose auxiliary word embeddings to improve the quality of topics. The effectiveness of our proposedmodel is comprehensively verified on three different types of datasets which are hyperlinked networks of web pages, social networks offriends and coauthor networks of publications. Experimental results show that our approach achieves significant improvementscompared with the current state of the art.

**Architecture**

****

**Existing System**

Mixture Membership Stochastic Blockmodel (MMSB) is one of the state-of-theart generative models, which can simultaneously factorize the network and infer the latent communities.LCTA (Latent Community Topic Analysis), one community can correspond to multiple topics and multiple communities can share the same topic. Consider that topics and communities are inner-dependent. Whether a link exists between two documents follows a binomial distribution parameterized by the similarity between topic mixtures and community mixtures. Our proposed model is more powerful in both topic modeling and community detection. Meanwhile, it is also elegant in capturing the semantic relations between topics and communities compared to previous work.

**Proposed System**

Proposed model LCTA (Latent Community Topic Analysis), one community can correspond to multiple topics and multiple communities can share the same topicalso consider that topics and communities are inner-dependent. Whether a link exists between two documents follows a binomial distribution parameterized by the similarity between topic mixtures and community mixtures. Our proposed model is more powerful in both topic modeling and community detection. Meanwhile, it is also elegant in capturing the semantic relations between topics and communities compared to previous work.

**Future Work**

The introduction of CCA makes our model capable to analyze correlations between the heterogeneous topics and communities, as well better model the heterophily (i.e., tendency to link to different others) and homophily (i.e., tendency to link to similar others) attributes of nodes. Comprehensive evaluations on three different datasets show that CTTA outperforms state-of-theart baselines with significant improvements.With increasing studies focusing on document networks, we believe that our proposed model is promising to advance the researches in this field. The application of CCA also offers inspirations for studies which would like to discover correlations between two sets of variables in other fields.

**Module Implementation**

1. **Canonical Correlation Analysis**

The introduction of Canonical Correlation Analysis (CCA) makes our model properly fit the dataset and need no prior hypothesis, such as nodes closely linked with each other share common topics. We derive an efficient mean-field variational EM algorithm for approximate posterior inference, which also significantly remedies the computational burden. Comprehensive evaluations on three different datasets are conducted to compare the proposed model with state-ofthe- art baselines. The results verify the excellent performance of CTTA.

1. **Mixed Membership Block Model**

Similar to most community detection algorithms, mixed membership block model (MMSB) summarizes a collection of pairwise measurements with a mapping from nodes to sets of nodes, called blocks, and pairwise relations among the blocks themselves.

1. **Word Embeddings**

Word embedding aims to learn high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. This section gives a very brief introduction about the Skip-gram model whose objective is to find word representations that are useful for predicting the surrounding words in a sentence or a document.

1. **Canonical Topical-Topological Analysis**

In this module, we present the details of our proposed Canonical Topical-Topological Analysis (CTTA) model, which simultaneously explores the topical and topological factors in document networks. For the latent topic factor, we appeal to CTM, which improves LDA by replacing the Dirichlet prior with normal distribution to capture the topical correlations. By coincidence, both random sets are drawn from normal distributions in the generative process of CCA. Thus we can naturally integrate CCA and CTM by fitting the per-document topic distribution prior. The topical correlation is explicitly captured by the covariance matrix of the normal distribution. For the latent community factor, we employ MMSB to factorize the network and infer latent clusters.

**Algorithm**

1. **Canonical Correlation Analysis (CCA)**

Which is a well-known machine learning algorithm is then applied to capture the latent semantic association between them. CCA aims to maximize the covariance between two random sets, thus it builds a reasonable bridge for topics and communities to interact with each other. This not only makes our model elegant in algebra but also totally unsupervised in implementation. On account of these advantages, we name our model Canonical Topical-Topological Analysis (CTTA). Specifically, it is a unified model which can simultaneously perform topic modeling and community detection, as well as discovering the correlations between them.

1. **Efficient Mean-Field Variational(EM)**

We derive an efficient mean-field variational EM algorithm for approximate posterior inference, which also significantly remedies the computational burden.

**SYSTEM REQUIREMENTS**

➢ **H/W System Configuration:-**

➢ Processor - Pentium –IV or Later Version

➢ RAM - 4 GB (min)

➢ Hard Disk - 40 GB

➢ Key Board - Standard Windows Keyboard

➢ Mouse - Two or Three Button Mouse

➢ Monitor - SVGA

**Software Requirements:**

* Operating System - Windows XP or Later Version
* Coding Language - Java/J2EE(JSP,Servlet)
* Front End - J2EE
* Back End - MySQL

**Conclusion**

This paper proposes a novel model CTTA which simultaneously performs topic modeling and community detection on document networks. The introduction of CCA makes our model capable to analyze correlations between the heterogeneous topics and communities, as well better model the heterophily (i.e., tendency to link to different others) and homophily (i.e., tendency to link to similar others) attributes of nodes. Comprehensive evaluations on three different datasets show that CTTA outperforms state-of-theart baselines with significant improvements.With increasing studies focusing on document networks, we believe that our proposed model is promising to advance the researches in this field. The application of CCA also offers inspirations for studies which would like to discover correlations between two sets of variables in other fields.