Correlated Topic Modeling for Big Data with MapReduce

Mi Khine Oo
University of Computer Studies, Yangon
Yangon, Myanmar
mikhineoo@ucsy.edu.mm

May Aye Khine
University of Computer Studies, Yangon
Yangon, Myanmar
mayayekhine@ucsy.edu.mm

***Abstract*—Efficient extraction of useful information is a rising problem in Big data, since the amount of information being gathered across various domains grows with an increasing rate. So, it takes more time to understand the underlying themes of the documents collection. To deal with such problem in the context of Big data, the proposed approach implements the correlated topic model (CTM) with MapReduce framework to reveal the thematic information represented by words, to speed up the processing and to increase the scalability of the model. We apply variational Expectation-Maximization (EM) to make inference for CTM. In this paper, academic articles are collected by using a web crawler. Then CTM is exploited to uncover the underlying themes of the collection. The use of CTM with MapReduce implementation improves the accuracy and performance in a reliable and scalable manner.**

Keywords—Big data, Correlated Topic model, Hadoop MapReduce Framework

I. INTRODUCTION

Among the topic modeling methods, the Latent Dirichlet Allocation (LDA) [1] is widely known and mostly applied to observe the knowledge from text collections. A limitation of LDA is the incapability of modeling topic correlations because LDA models the topic proportions by using a Dirichlet distribution. The Correlated Topic Model (CTM), which was first proposed in [2], figures out this limitation by replacing the Dirichlet distribution with the logistic normal distribution to reveal the correlations of latent topics.

In order to extract the latent topics, different inference methods have been proposed to estimate the model parameters, such as Gibbs Sampling and Variational Expectation-Maximization (EM). The purpose of this paper is to develop a scalable CTM with variational EM algorithm to process Big data in a Hadoop cluster.

The previous studies in the field of topic modeling utilized open-source frameworks to handle Big data. Zhai et al. [3] developed a parallelized Mr.LDA algorithm using variational inference based on the MapReduce. However, the extracted topics can correlate with each other in realistic applications. Aznag et al. [4] applied CTM to discover and rank web services descriptions. Sang et al. [5] also used CTM to recognize facial expressions with two types of facial features.

Extracting meaningful topics from academic documents collection in Big data is still an ongoing research. A summary expression of the major contributions is as follows:

1. Proposing a Big data CTM model to extract the latent topics by the use of variational EM with MapReduce.
2. Searching and collecting full-text research documents from various domains by developing a web crawler. We considered to process the full-text academic papers in order to increase the accuracy of the extracted topics.
3. Implementing the model with MapReduce framework to enhance its capacity for Big data.

The structure of this paper has been organized with different sections. Section II presents the theory background of CTM. The workflow of proposed approach is described in section III. Finally, section IV focus on conclusion and further extensions of this study.

II. CORRELATED TOPIC MODEL

CTM [2] models that documents are mixtures of a set of hidden topics, where each topic is drawn from a distribution over words, which are all present in a vocabulary. Fig. 1 illustrates the graphical model representation for CTM. The notations which are used in this paper related with CTM are described in TABLE 1.

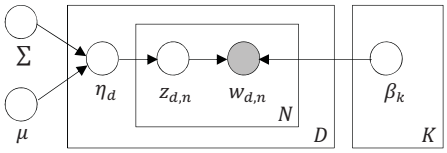

Fig. 1. Graphical model of CTM

Given a documents collection D, number of topics K, and K-dimensional mean and K×K covariance matrix $\{\mu, \Sigma\}$, the generative process of CTM is as follows [6]:

1. For each topic in $k \in K$
   Draw a multinomial $\beta_k \sim N(\mu, \Sigma)$ over vocabulary
2. For each document $d \in D$
   a. Draw a topic distribution $\eta_d \sim N(\mu, \Sigma)$ over all topics
   b. For each word $n \in \{1, \ldots, N_d\}$
      i. Choose a topic assignment $z_{d,n} \sim Mult(\theta)$
      ii. Choose a word $w_{d,n} \sim Mult(\beta_{z_{d,n}})$

TABLE I. NOTATIONS

Symbol	Description
D	Total documents in collection
N_d	Total words in document *d*
V	Total words in vocabulary
K	Number of latent topics
β_k	Word distribution per topic
η_d	Topic distribution per document
θ	Document-specific topic proportions
$\mu, \sum$	Logistic normal parameters
$z_{d,n}$	Topic assignment for word *n* in document *d*
$w_{d,n}$	Word assignment for word *n* in document *d*

Given a document w and a model $\{\beta, \mu, \Sigma\}$, the posterior distribution of the latent variables $p(\eta, z \mid w, \beta, \mu, \Sigma)$ is intractable to compute. Use Jensen's inequality to bound the log probability of a document,

$$\log p(w_{1:N} \mid \mu, \Sigma, \beta) \geq E_q[\log p(\eta \mid \mu, \Sigma)] + \sum_{n=1}^{N} E_q[\log p(z_n \mid \eta)] + \sum_{n=1}^{N} E_q[\log p(w_n \mid z_n, \beta)] + H(q) \quad (1)$$

where $H(q)$ denotes the entropy of variational distribution. To describe q, a factorized distribution parameterized by the variational parameters is used,

$$q(\eta, z \mid \lambda, \nu^2, \phi) = \prod_{i=1}^{K} q(\eta_i \mid \lambda_i, \nu_i^2) \prod_{n=1}^{N} q(z_n \mid \phi_n) \quad (2)$$

The variational inference optimizes Equation (1) by updating the variational parameters to minimize the Kullback-Leibler (KL) divergence between q and the true posterior p.

After the variational inference, a variational expectation-maximization (EM) is applied to maximize the likelihood of documents collection. A variational inference for each document is done to maximize the bound with respect to the variational and model parameters in the E-step and M-step respectively.

III. PROPOSED APPROACH

In this section, we propose to develop a CTM with variational EM algorithm in MapReduce framework to improve the scalability for Big data.

A. Data Collection

The system employs a web crawler to gather the full-text documents from the publicly available digital libraries: CiteSeerX, JSTOR and PLOS ONE. The crawler first visits these webpages and finds the PDF documents because most of the academic articles are only published in PDF formats. Each crawled document is converted into its textual format (.txt) by using the Apache PDFBox text conversion tool. The resulting text files are uploaded to HDFS to perform preprocessing.

B. Data Preprocessing

The input is a crawled text documents collection from a large number of domains, which is called the Big data. The *Map* function splits each line of the text document into words and emits a <key, value> pair for each word, where the key is the word and the value is 1. The numbers, special characters, stopwords and short terms with less than four characters are removed. The most frequent terms and terms appearing less than five times are also eliminated during in the Map phase. The *Reduce* function produces the list of words with the count of occurrence of each word.

C. Model Training

For training CTM, we adopt the variational EM of CTM as described in section II. The Driver program takes the control of the whole process and implements the MapReduce jobs. The Mapper algorithm performs the E-step and the Reducer algorithm performs the M-step.

Algorithm 1 MapReduce Driver for CTM

Input: Number of topics K, A corpus consisting of D documents, N_d words in document d
Output: Model parameters $\{\beta_{1:K}, \mu, \Sigma\}$
Initialize variational parameters: $\lambda_i = 0, \upsilon_i^2 = 0, \zeta = 0, \phi_{n,i} = 1/K$ for all i in K and n in N_d
Initialize model parameters: $\mu = 0, \Sigma = 1, \beta_i = 0.01 + \text{rand}()$
  Call Mapper algorithm
  Call Reducer algorithm

Algorithm 2 Mapper for CTM

repeat
  for $d = 1$ to D
    for $n = 1$ to N_d
      for $i = 1$ to K
        Update ζ with $\zeta = \sum_i \exp(\lambda_i + \upsilon_i^2/2)$
        Update $\phi_{n,i}$ with $\phi_{n,i} = \exp(\lambda_i)\ \beta_{i,n}$
      end for
    end for
    Update $\phi_{d,i} = \sum_n \phi_{n,i}$
    Update λ_i with $dL/d\lambda = -\sum^{-1}(\lambda - \mu) + \sum_{n=1}^{N} \phi_{n,1:K} - (N/\zeta)\ \exp\{\lambda + \upsilon^2/2\}$
    Update υ_i^2 with $dL/d\upsilon_i^2 = -\sum_{ii}^{-1}/2 - N/2\zeta \exp\{\lambda + \upsilon_i^2/2\} + 1/(2\upsilon_i^2)$
  end for
until convergence
Emit results to Reducer

Algorithm 3 Reducer for CTM

for $d = 1$ to D
  for $i = 1$ to K
    Update β_i with $\beta_i = \sum_d \phi_{d,i}\, n_d$
  end for
  Update μ with $\mu = \frac{1}{D}\ \sum_d \lambda_d$
  Update Σ with $\Sigma = \frac{1}{D}\ (\sum_d diag(\upsilon_d^2) + \sum_d (\lambda_d - \mu)(\lambda_d - \mu)^T)$
end for
Emit results to MapReduce Driver

IV. CONCLUSION

Correlated topic modeling is a powerful technique for extracting value from Big data. In this paper, CTM with the variational EM algorithm for Big data is proposed to extract the topics that represent the corpus. Our proposed approach is still in the developing stage and intends to increase the quality of the latent topics, to reveal the inter-related topics of the corpus, and to improve the reliability and scalability of CTM by the use of variational EM over MapReduce framework. In the future, our research will draw attention to improve the performance of the variational EM algorithm. We will also develop an application using CTM based on the MapReduce framework.

REFERENCES

[1] M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation," The Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2013.

[2] M. Blei and J. D. Lafferty, "Correlated topic models," Advances in Neural Information Processing Systems, MIT Press, Cambridge, MA, 2006.

[3] K. Zhai, J.L. Boyd-Graber, N. Asadi, and M.L. Alkhouja, "Mr. LDA: a flexible large-scale topic modeling package using variational inference in MapReduce," WWW, pp. 879–888, 2012.

[4] M. Aznag, M. Quafafou and Z. Jarir, "Correlated topic model for web services ranking," IJACSA, 4(6), 2013.

[5] R. Sang and K. P. Chan, "A correlated topic modeling approach for facial expression recognition," IEEE (CIT2015), 2015.

[6] M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, Vol. 1, No. 1, pp. 17–35, 2007.