Project proposal

Hawkes process: Information diffusion and Topic modeling from text-based cascades.

Yutong Wang
yw4ku@virginia.edu

Yiming Wang
yw5hn@virginia.edu

Yingqing Huang
yh3ra@virginia.edu

ABSTRACT

In this work, we are interested in (1) investigating the hidden influence network in a collection of documents, given document content and timestamps, and (2) find the appropriate set of "influence nodes" for a specific document *D*, where "influence nodes" can be think of the nodes that caused document *D* to be published. We propose to solve the first problem by using Hawkes process, and solve the second one using Topic Correlation Model.

KEYWORDS

Hawkes process, topic modeling, information diffusion

1 INTRODUCTION

Today in scientific community, a large number of research papers are published every year. In many areas, it is often the case that one research paper is based on another (or several) previous related research, with improvements in some aspects. It is therefore sometimes difficult for a person new to a field to find the appropriate research paper to start with.

In this work, we propose to solve this problem by recommending the past research that user is interested in. The solution is divided in two steps: first, investigate the diffusion process of the documents using Hawkes Process; second, represent each document as a set of abstract topics using Correlated Topic Model. Solving the first problem allow us to deduce a hidden influence network (i.e. which paper influenced which), and all the nodes influenced user's matched result. Solving the second one allows us to better understand the content of the document and infer which parent node to choose from.

2 MANAGEMENT PLAN AND DETAILS

2.1 Management plan:

1. Read and understand past research papers.
2. Implement Hawkes Process (1) and Topic Correlation model (1).
3. Test the program on real world data and collect result (1).
4. Finish the report.

The specific assignment of task for everyone is subject to change based on difficulty of tasks. If we have extra time left, we will also work on combining the two models.

2.2 Details

Hawkes process: Hawkes Process has been extensively used in financial analysis and earthquake modeling due to its self-exciting property, that is, each event happened increases the rate of future event happening for some period of time [2]. We believe it is a perfect model for inferring the hidden influence network in our study, in that one publish (e.g. research, news) could trigger a series of future related publishes. This influence then gradually dies out as time passes by (More and more paper base on new research; news become obsolete in a few days).

LDA: One common technique used in topic modeling is La tent Dirichlet Allocation (LDA) [5]. However, LDA suffers from the inability to model topic correlation. Therefore, to solve the first problem, we decided to use Correlated Topic Model (CTM) [3], which captures the topics of each docum ent more accurately. For the second problem, i.e. understan d the diffusion process, we use Hawkes-Process (HP) as ou r model due to its well-known self-exciting property. Figur e 1. explains the process of LDA. This process is identical t o the generative process of LDA except that the topic propo rtions are drawn from a logistic normal rather than a Dirichl et.

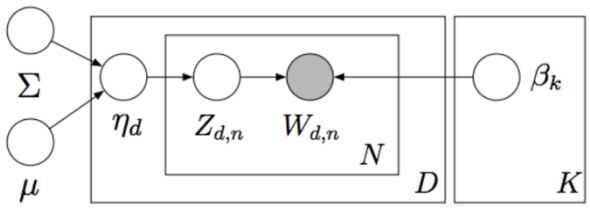

Figure 1. Graphic representation of Correlated topic model

Note that although our model is built for research papers or news articles, the methodology of our model goes beyond that. For example, one can use our model to detect how a certain tweet got spread over time by adjusting topic modeling to be suitable for documents with a small number of words.

3 DATASET

We want to choose data that have hidden influence over eac h other, so that HP can be applied in modeling. We also wa nt the number of words in each document is relatively larg e, so that the CTM can give a more accurate result. One lik ely choice, for example, is a set of academic papers in a cer tain field.

REFERENCES

[1] Xinran He, Theodoros Rekatsinas, James Foulds, Lise Getoor and Yan Liu, HawkesTopic: A Joint Model for Network Inference and Topic Modeling from Text-Based Cascades, ICML'15.

[2] Liniger, Thomas Josef. Multivariate Hawkes processes. PhD thesis, ETH Zurich University, 2009.

[3] David M. Blei and John D. Lafferty. A Correlated Topic Model of Science. The Annals of Applied Statistics Vol. 1, No. 1 (Jun., 2007), pp. 17-35.

[4] Aleksandr Simma , Michael I. Jordan, Modeling events in time using cascades of poisson processes, University of California at Berkeley, Berkeley, CA, 2010.

[5] Latent Dirichlet allocation: https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation