Methods for News Items Popularity Estimation on Early Stages

Avetisyan A.[1,3] Drobyshevskiy M.[1,2] Turdakov D.[1,3]

[1]Ivannikov Institute for System Programming of the RAS

[2]Moscow Institute of Physics and Technology, Moscow, Russia

[3]Lomonosov Moscow State University, Moscow, Russia

ISPRAS OPEN, 2019

Introduction

- Prediction of the popularity of data streams can be used in various scenarios, such as political campaigns, preventing the spread of fake news
- Most of news platforms do not have explicit links between each other.

Directions of studying information propagation

- Inferring an influence graph under the assumption that information is spread along its edges
- Prediction of news popularity based on information flow features (temporal, structural, content, features of early adopters)

Definitions

- **information propagation for a particular message (cascade)** is a publication of news item on the network by any user (post):

  $$c = ((u_1, t_1, \theta_1), (u_2, t_2, \theta_2), ..., (u_n, t_n, \theta_n)),$$

  where u_i is the i-th disseminator who posted the message, t_i is a corresponding publication time and θ_i is an information about the post.
- A cascade will be considered **popular** if it contains more messages than $n\%$ of other cascades in the flow. We will consider $n = 50\%$ in this paper.
- **The early stage of a cascade** is the time when a small fixed number k of disseminators posted the message. We will consider $k = 5$ in this paper.

Formulation of the problem

Input data

- set of cascades C, each of which is a sequence of publications of one post represented in the chain

$$c = ((u_1, t_1, \theta_1), (u_2, t_2, \theta_2), ..., (u_n, t_n, \theta_n)),$$

where u_i is the i-th disseminator who posted the message, t_i is a corresponding publication time and θ_i is an information about the post.

Task

To predict for every cascade c in the flow of propagated messages at an early stage (k = 5) whether $|c|$ will be greater than $n\%$ of other cascades in the flow over time t

Structural features

Assumption

There is a hidden graph of influence between network nodes. Vertices are sources of information, while the edge (u, v) means that the source u affects v.

Social network graphs

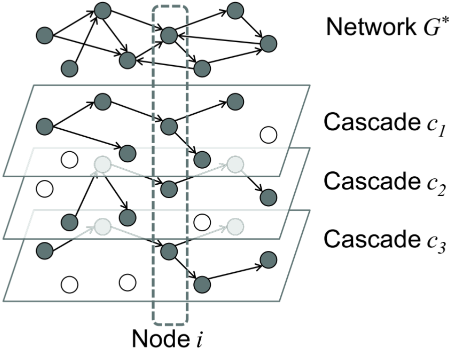
For networks with a social graph, consider that this social graph is a graph of influence

Proposed Model for News Items

In the absence of a social structure, we build a hidden graph of influence using NetInf algorithm[a] based on a given set of cascades.

[a] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks ofdiffusion and influence, 2010

NetInf



- NetInf predicts the most likely hidden graph of influence based on a set of cascades
- NetInf is an iterative greedy algorithm which maximises the likelihood function
- At each iteration NetInf finds the most influential edge and adds it to the graph of influence

Method

Features

- Temporal features (time intervals between early adopters)
- Structural features (early adopters degrees)
- Features of early adopters (average number of publications per day)
- Content features (news topic[a], text similarity)

[a] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation, 2003

Classifier

XGBoost classifier for predicting, where parameters *max_depth* and *min_child_weight* were tuned using cross-validation. Other parameters were taken by default.

Experimental evaluation

Data

- Lastfm[a] is a music website which have a social graph where users can listen to the music and mark the songs they like. 212 000 cascades for 450 000 users.
- Yandex data: 68 000 cascades for 2500 news publications at Yandex news service from January 2016.

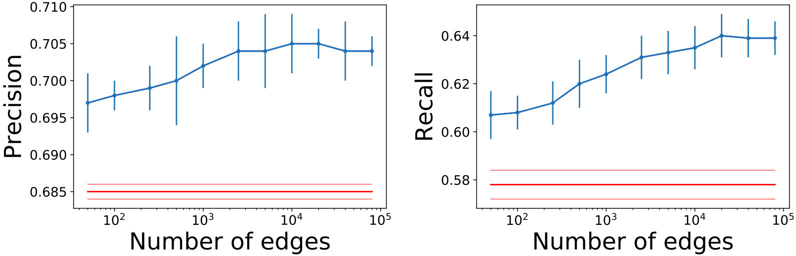
[a] B. Shulman, A. Sharma, and D. Cosley. Predictability of popularity: gaps between prediction and understanding, 2016

Two stages

1) Choosing the best hidden graph of influence for prediction

2) Predicting cascade popularity using different types of features.Two metrics were used for evaluation of the model: precision and recall.

Experiments 1



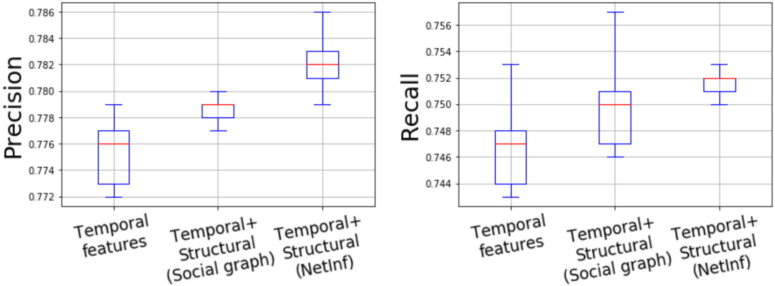
Precision and recall for the predicting model using temporal (red) and temporal+structural (blue) features extracted from the hidden graph of influence with different number of edges using Yandex cascades.

Experiments 2

Table: Yandex cascade prediction. Structural features are obtained from the hidden graph of influence with k = 20000 edges

Types of features	***Precision***	***Recall***
Temporal	0.685 ± 0.001	0.578 ± 0.006
Temporal + Structural	0.705 ± 0.002	0.640 ± 0.009
Temporal + Structural + Early Adopters	0.736 ± 0.002	0.675 ± 0.011
Temporal + Structural + Early Adopters + Content	**0.750** ± 0.004	**0.722** ± 0.008

Experiments 3



Precision and recall for the predicting model using temporal, temporal+structural (social graph) and temporal+structural (NetInf) features with 100000 edges in the hidden graph using Lastfm cascades.

Types of features	***Precision***	***Recall***
Temporal	0.776	0.747
Temporal + Structural (Social graph)	0.778	0.750
Temporal + Structural (NetInf)	0.782	0.752

Discussion

- We proposed a model which predicts news stories popularity at the early stage and reconstructs a graph of influence in the absence of the social graph which improves the prediction quality.
- Structural features based on a constructed graph improves the prediction. precision and recall for Yandex cascades by 2% and 6%, respectively.
- Using of all four types of features (temporal, structural, early adopters, and content) significantly improves the model for Yandex data compared to the use of only temporal features. Precision and recall improve by 7% and 15%, respectively.
- Using the NetInf algorithm allowed to achieve similar or even better prediction quality than using the original social graph.