User community detection via embedding of social network structure and temporal content

By Fani et al.

EGOR DMITRIEV, Utrecht Univeristy, The Netherlands

1 GOALS

- Explore various aspects related to the identification of user communities
- we identify user communities through multimodal feature learning (embeddings)
  - we propose a new method for learning neural embeddings for users based on their **temporal content similarity**
  - we learn user embeddings based on their **social network connections**
  - we systematically **interpolate** temporal content-based embeddings and social link-based embeddings
- we systematically evaluate the quality of each embedding type in isolation
- Application scenarios, namely news recommendation and user prediction

2 PRELIMINARIES

- Random Walk:
  - BFS favours structural equivalence
  - DFS in contrast, respects homophily and leads to similar (close) embeddings for densely connected users

3 CHALLENGES

- Given content on the social network are often reflective of issues in the real world, the **topics discussed on the network constantly change** and hence users' interests towards these topics also change as the community evolves and new topics and connections are made
- Users are essentially interested in the same topic but their interests are temporally distributed differently over time

4 PREVIOUS WORK / CITATIONS

- Content Based Methods:
  - Modeled communities **based on topics of interest** through a community-user-topic generative process
  - **Communities** are **formed around multiple correlated topics** where each topic can be reused in several different communities
  - User can be a member of different communities but with varying degrees of membership
  - **Distributional semantics** states that words that occur in similar contexts are semantically similar
- Link-based methods:
  - Primarily based on the **homophily** principle (densely connected groups of users imply a user community)
  - Explicit social connection does not necessarily indicate user interest similarity, but could be owing to **sociological processes**

- **This Work:** ...

5 DEFINITIONS

- Those users who share not only similar topical interests but also share similar temporal behavior are considered to be like-minded and hence members of the same community
- Temporal Social Content: $\mathcal{D} = (\mathbb{U}, \mathbb{M}, \mathrm{T})$
  - $\mathbb{U}$: Users, $\mathbb{M}$: Text content, T time periods
- Social Network Graph: $\mathcal{G} = (\mathbb{U}, \mathbb{A})$
  - $\mathbb{U}$: Users/Nodes, $\mathbb{A}$: Edges

6 OUTLINE / STRUCTURE

- Problem: Cluster Users into groups where users are more similar compared to intra group similarity

6.1 Temporal content-based user embeddings:

- Identify a set of topics $\mathbb{Z}$ from $\mathbb{M}$
- Construct **user's topic preference timeseries**: $\mathrm{X}_{uz} = \left[x_{uz,1} \cdot x_{uz,\,\mathrm{T}}\right]$
- $x_{uz,t} \in \mathbb{R}^{[0,1]}$
- Each user needs to be defined in the context of other users
  - The more two users share common interests in similar time intervals, the more similar these users would be and hence the likelihood of these users being in the same community should increase
- **Region of like-mindedness**: Parts in X where users share interest in same topics given a threshold (for level of interest)
- Adopt the **continuous bag-of-word** (CBOW)
  - Context for each user to consist of all those users who have been observed with this user in similar region's of like-mindedness
  - Use a one-hot encoding representation to refer to users in the input and output layers
  - Use a hierarchical softmax

6.2 Neighborhood context model

- Predict observing users such as u from v's neighborhood, adopting **skip-gram** model from
  - Again predict which users correspond (softmax)
- Sample using random walk

6.3 Embeddings interpolation

- $h\left(\mathrm{W}_{\mathcal{D}}, \mathrm{W}_{\mathcal{G}}\right) = \alpha \mathrm{W}_{\mathcal{D}} + (1-\alpha)\mathrm{W}_{\mathcal{G}}$
- Simply weigh the two embeddings

6.4 Community detection

- Identify communities of users through graph-based partitioning heuristics
- Construct a weighted graph: $\mathrm{G} = (\mathbb{U}, \mathbb{E}, w)$
  - With as weights the user embedding dot products
- Leverage the Louvain Method (LM)
  - Can be applied to weighted graphs
  - Does not require a priori knowledge of the number of partitions
  - Has an efficient linear time complexity

7 EVALUATION

- Content-based methods produce higher quality communities compared to link-based methods
- Methods that consider temporal evolution of content, our proposed method in particular, show better performance
- Communities that are produced when time is explicitly incorporated in user vector representations have higher quality
- Our experiments show that while social network structure is not a discriminative enough feature on its own for identifying high quality user communities, it does improve the quality of the identified user communities when effectively interpolated with content-based methods.

8 CODE

- ...

9 RESOURCES

- ...

10 NOTES:

- No negative sampling (for user embedding learning by context?)