Inferring Shared Interests from Social Networks

Laura Dietz*
Max-Planck Institute for Computer Science, Saarbrucken, Germany
dietz@mpi-inf.mpg.de

Abstract

We present two extensions of latent Dirichlet allocation: the citation influence model and the shared taste model. Both aim towards extracting shared interests and quantifying the mutual influence on each other. The extraction is unsupervised, the only source being a social network with additional user data or a citation network with abstracts.

1 Introduction

Many online community platforms such as LibraryThing, last.fm, flickr, and CiteULike store data about users, friend relationships between users, and for each user a list of items he interacted with. Depending on the usage scenario, items may be books, songs, pictures, or scientific publications, respectively. In social network analysis it is widely assumed that people tend to gather in groups of shared interests, where such interests drive friendships and vice versa. The same assumption is made in citation networks. Citation networks are directed and acyclic, where social networks are symmetric and have small diameter.

Typically, shared interests in networks are either not available at all or only on a very coarse granularity, such as conference venues and affiliations. The inference of shared interests is further complicated if we expect the network to arise from many overlapping interest groups with mixed membership.

Problem statement. This work focuses on unsupervised inference of shared interests—tastes or topics—$\mathcal{T}$ given a social network. Nodes $n \in \mathcal{N}$ in the network refer to users and edges $\mathcal{E}$ refer to online friendships between users. Each node has a set of associated items $\mathcal{C}(n)$ from a common item vocabulary $\mathcal{V}$. Given the network structure $(\mathcal{N}, \mathcal{E})$ and node contents $\mathcal{C}$, the goal is a) to learn shared topics $\mathcal{T}$ that follow human intuition and b) to quantify the mutual influence between nodes.

Applications. If the shared taste and influence would be known by the platform software, this would give rise to new social networking features. For citation networks, this allows to filter the citation network according to a specific research topic. Tools such as Google Scholar and CiteSeer, that allow to navigate in the publication graph provide a great support for researchers who need to get a quick overview about a research area. Inspired by the work of Garfield [1], we would like to create a bird's-eye visualization of a research area that complements in-depth navigation in the publication graph. But because the publication graph is linked densely, even a radius of two citations from a pivotal paper contains hundreds of publications. When examining the citations in detail, one finds that not all cited work has a significant impact on a citing publication. A bird's-eye visualization should show papers that significantly impact one another. This requires to measure the strength of a citation's influence on the citing work.

For social networking platforms that allow users to subscribe to items of their friends, this subscription feature can be improved. Typically each user has diverse interests, e.g., likes rock and jazz.

*Laura Dietz is supported by a scholarship of Microsoft Research Cambridge.

Algorithm 1 Generative process of the citation influence model (left) and shared taste model (right).

for all $t \in \mathcal{T}$ **do**
  draw $\phi_t \sim \text{Dirichlet}(\alpha_\phi)$
for all $c \in \mathcal{N}$ as cited role **do**
  draw $\theta_c \sim \text{Dir}(\alpha_\theta)$
  for all $x'_{c,i} \in \mathcal{C}(c)$ **do**
    draw $t'_{c,i} \sim \text{Multi}(\theta_c)$
    draw $x'_{c,i} \sim \text{Multi}(\phi_{t'_{c,i}})$
for all $d \in \mathcal{N}$ as citing role **do**
  draw $\gamma_d \sim \text{Dir}(\alpha_\gamma)$
  for all $x_{n,i} \in \mathcal{C}(n)$ **do**
    draw $c_{d,i} \sim \text{Multi}(\gamma_d)$
    draw $t_{d,i} \sim \text{Multi}(\theta_{c_{d,i}})$
    draw $x_{d,i} \sim \text{Multi}(\phi_{t_{d,i}})$

for all $t \in \mathcal{T}$ **do**
  draw $\phi_t \sim \text{Dir}(\alpha_\phi)$
for all $\{u, f\} \in \mathcal{E}$ **do**
  draw $\theta_{\{u,f\}} \sim \text{Dir}(\alpha_\theta)$
for all $u \in \mathcal{N}$ **do**
  draw $\psi_u \sim \text{Dir}(\alpha_\psi)$
  for all $x_{u,i} \in \mathcal{C}(n)$ **do**
    draw $f_{u,i} \sim \text{Multi}(\psi_u)$
    draw $t_{u,i} \sim \text{Multi}(\theta_{\{u,f_{u,i}\}})$
    draw $x_{u,i} \sim \text{Multi}(\phi_{t_{u,i}})$

Say, a subscribing users likes only jazz and classical music. A straight forward implementation of a subscription feature may annoy the subscribing user with rock songs, although the shared taste is clearly jazz. An improved implementation would re-weight the subscribed item list to match the shared taste. Further, knowing a set of typically shared interests allow for visualizing the social neighborhood of selected users as well as the citational neighborhood of selected papers.

Latent Dirichlet allocation (LDA) [2] allows to extract common topics from a corpus of documents ignoring any underlying graph structure. In this work, we study two extensions towards LDA where topic inference is guided by the graph structure, yielding different topics. The Pairwise Link-LDA model [3] explains the network with a stochastic mixed-membership blockmodel combined with LDA to explain contents. An underlying assumption is that absent edges indicate incompatible topics, an assumption that does not hold for networks in our case study.

2 Citation Influence Model

The citation influence model [4] explains strong influence between nodes if their content is compatible, i.e., parts of the content may have been taken from the neighboring node.

Citation networks. In citation networks, each publication represents a node n in the network, with words in the publication representing the contents $\mathcal{C}(n)$. Edges represent citation links from the citing towards the cited publication.

The approach devises a probabilistic model that explains the generation of documents d and their cited publications c. The intuition behind the model is that parts of the content in a citing publication are taken from its citations. It is assumed that the more text is associated to one citation, the more compatible the publications are and thus, the stronger is the influence it has on the citing publication. The relative influence of citations $c_1, c_2, ...$ on d is quantified by a multinomial parameter $\gamma_d(c)$. Rather than copying text on a word basis, the model incorporates a nested generative process, where words represent topics and cited publications are modeled by topic mixtures just like in latent Dirichlet allocation. The generative process is given in Algorithm 1 (left).

Robustness may be improved if we allow the model to leave some passages unassociated to any cites, as such passages represent innovational aspects either as own topics or own items. Topic mixtures θ, word distributions ϕ, and strength of influence γ are estimated conjointly to foster mutual benefit.

The generative process assumes that the contents of cited and citing publications are independent given the topic mixture. However, during parameter estimation, the topic mixture is influenced from all associated words, that is all words of the cited publication c together with parts of contents in linked publications $d_1, d_2, ...$ that cite c. As a result, topic mixtures are not only describing one publication, but the contents all neighbored publications that are drawn from each topic mixture. As an implication, the estimated word/item distribution ϕ_t will account for evolving vocabulary, for instance by associating the term "topic model" with the publication "Latent Dirichlet allocation."

The citation influence model will explain each publication twice: once for its role as a cited publication, and once as a citing publication. This leads to a duplication of nodes in the network towards a bipartite graph. As duplication yields an extreme co-occurrence pattern, and LDA assigns same topics to co-occurring items, grandparents will influence the topics of grandchildren via duplicates.

Social networks. A further asset of the bipartite transformation is that it allows to apply the model on social network data. Nodes are duplicated for the "friend" role f in analogy to cited publications and the "user" role u in analogy to the citing publications d. The estimated compatibility $\gamma_u(f)$ indicates the influence that the friend f has on u's contents. As symmetric edges of social networks are treated as bi-directional edges in the model, estimated topic mixtures represent shared interests of a node-neighborhood.

3 Shared Taste Model

The node duplication of the citation influence model is not an elegant solution. The shared taste model builds on the same assumptions, but yields a more natural model for learning shared topics and quantifies the influence. In contrast to having topic mixtures represent node-neighborhoods, the shared taste model [5] uses topic mixtures to explain friendships, exclusively modeling contents of the two incident nodes.

Social networks. The shared taste model exploits the symmetric character of friendship by associating each edge with a shared topic mixture θ. Each user is associated with a mixture ψ over friends or incident edges, quantifying influence in relation to other friends of the user. Each item of a node's content is explained by one of the node's friendships in drawing a topic from the common edge's mixture. The item is explained by the topic's item distribution ϕ following latent Dirichlet allocation. The full generative process is given in Algorithm 1 (right). As with the citation influence model, robustness can be improved by leaving some items unassigned to any friendship. Topic mixtures θ and item distribution ϕ are estimated together with the influence ψ.

During inference, the shared topic mixture is influenced by both users' contents to a degree that varies with the influence ψ. As the contents of two friends u and f are explained by a shared topic mixture (associated with the edge $\{u, f\}$), the model is encouraged to use topics that match both users' contents well. This introduces a coupling of topics across the network. Despite the fact that estimated topic mixtures θ refer to edges instead of users, node-centered topic mixtures can be obtained post inference by aggregating topics from all incident edges weighed by the influence ψ.

Citation networks. Ignoring the direction of edges, the shared taste model is readily applied to citation networks, learning link-based topics and node-influences in a unified manner.

4 Experiments

LibraryThing[1] is a social networking platform centered around books. The platform hosts user groups, providing a group-wise discussion forum and suggested reads centered around a common subject. We study whether the two models identify the groups a user is member of from the tags she used for her library and friend relationships in an unsupervised manner. Thus, nodes $\mathcal{N}$ represent users, edges $\mathcal{E}$ represent online friendships and items $\mathcal{V}$ refer to tags. We selected a connected subgraph with 194 users, a tag vocabulary of size 748, and 10 user groups. Groups are not part of the training data—they only serve as a ground truth for evaluation. As we are interested in finding the original ten user groups, we set the number of different topics $|\mathcal{T}|$ to 10. The inference for all models is implemented using Infer.NET.[2]

We study the correlation between predicted topics and the true membership in user groups by Pearson correlation coefficient (Figure 1 left and center). As the correlation matrices are hard to compare quantatively, we summarize each by one scalar number that expresses the consistency of topic-group correlation. For that we use the correlation coefficient to predict group-membership for each user. Results are given as averaged AUC values with standard error bars in Figure 1 right.

[1] http://www.librarything.com

[2] http://research.microsoft.com/infernet

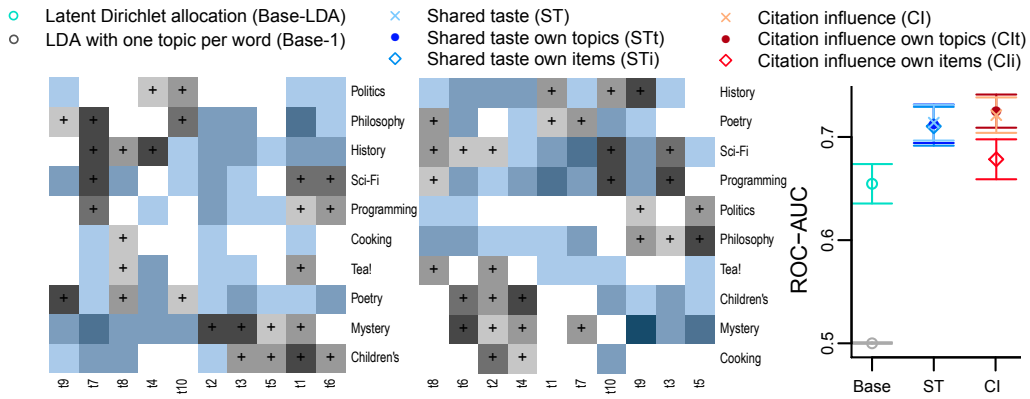

Figure 1: Pearson correlations between topics and held-out groups on LibraryThing. Left: shared taste model; Center: citation influence model. (Black/+: positive correlation; blue: negative correlation; white: uncorrelated.) Right: Summary of correlation consistency, higher is better. Categories in x-axis matches the column in the legend.

Both the shared taste and the citation influence model infer reasonable topics, with no significant difference found by a paired-t-test with 5% level. From the raw data analysis we know that programmers like sci-fi. Both models rediscover this, manifested in topics t6 and t7 for the shared taste model, and in topics t10 and t3 for the citation influence model. Analyzing the words assigned with topics, we found that the shared taste model is slightly better in distinguishing between sci-fi related terms (t7) and computer-related terms (t6). Both models significantly outperform LDA.

5 Conclusions

We presented the citation influence model and the shared taste model for social networks. Both extend latent Dirichlet allocation towards exploiting the graph structure in order to learn shared interests. Both models have a notion of varying influences between linked nodes and shared topic mixtures. The topics indicate the topical role that nodes play in their network vicinity. The shared taste model identifies the common taste of each friendship and thus yields slightly more fine-grained topics, where the citation influence model learns topics shared by a neighborhood of nodes.

Although both models can be applied to citation and social networks, the shared taste model does not require nodes to be duplicated. Thus, it is a natural choice for undirected networks with similar modeling assumptions and prediction performance as the citation influence model.

The shared interests give rise to new visualizations and new features for social networking platforms. With the wide-spread use of web 2.0, the internet is becoming one large social network where ties such as friendships, one-way subscriptions, and web-page visits are formed according to shared interests. In a society that suffers from information overload, inference of and filtering using shared interests will help people to focus on information the are ultimately interested in.

References

[1] Eugene Garfield. Historiographic mapping of knowledge domains literature. *Journal of Information Science*, 30(2):119–145, 2004.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William. Joint latent topic models for text and citations. In *International Conference on Knowledge Discovery and Data Mining*, 2008.

[4] Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, 2007.

[5] Laura Dietz. Modeling shared tastes in online communities. In *NIPS Workshop on Applications for Topic Models: Text and Beyond*, 2009.