Extracting Community Structure through Relational Hypergraphs

Yu-Ru Lin[1] Jimeng Sun[2] Paul Castro[2] Ravi Konuru[2] Hari Sundaram[1] Aisling Kelliher[1]

[1]Arts Media and Engineering
Program, Arizona State University,
Tempe, AZ 85281 USA

[2]IBM T.J. Watson Research Center,
Hawthorne, NY 10532 USA

{yu-ru.lin, hari.sundaram, aisling.kelliher}@asu.edu, {jimeng, castrop,rkonuru}@us.ibm.com

ABSTRACT

Social media websites promote diverse user interaction on media objects as well as user actions with respect to other users. The goal of this work is to discover community structure in rich media social networks, and observe how it evolves over time, through analysis of multi-relational data. The problem is important in the enterprise domain where extracting emergent community structure on enterprise social media, can help in forming new collaborative teams, aid in expertise discovery, and guide long term enterprise reorganization. Our approach consists of three main parts: (1) a relational hypergraph model for modeling various social context and interactions; (2) a novel hypergraph factorization method for community extraction on multi-relational social data; (3) an on-line method to handle temporal evolution through incremental hypergraph factorization. Extensive experiments on real-world enterprise data suggest that our technique is scalable and can extract meaningful communities. To evaluate the quality of our mining results, we use our method to predict users' future interests. Our prediction outperforms baseline methods (frequency counts, pLSA) by 36-250% on the average, indicating the utility of leveraging multi-relational social context by using our method.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—

Data mining; H.3.3 [**Information Storage and Retrieval**]:

Information Search and Retrieval—*Information filtering*

General Terms

Algorithms, Experimentation, Measurement, Theory

Keywords

Community Evolution, Relational Hypergraph, Non-negative Tensor Factorization, Dynamic Social Network Analysis

1. INTRODUCTION

Today, users routinely produce and consume media as well as interact with each other on social networking websites (e.g. Flickr, facebook). These sites allow a wide array of actions on media objects (e.g. uploading photos and bookmarking), as well as actions with respect to other users (e.g. instant messaging). The interaction among users can be explicit (e.g. via instant messaging), or implicit (two users may share similar tags, or read a common post). Enterprises have increasingly embraced social media software to promote collaboration. Such social media (Wikis, bookmark sharing, etc.) foster dynamic collaboration patterns that deviate from the formal organizational structure (e.g. cooperate departments, etc.). The primary motivation for this work is to extract emergent community structure in the enterprise through the analysis of user interaction over social media. This will have significant impact – it can help foster new collaborative teams, help with expertise discovery and in the long term, guide enterprise reorganization consistent with collaboration patterns.

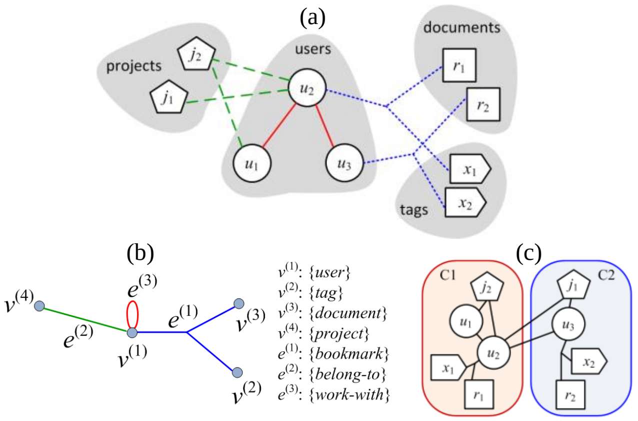


Figure 1: (a) an example of facets and relations in an collaborative environment; (b) a relational hypergraph for representing (a); (c) implicit community structure in (a) – C1 and C2 are two clusters we want to identify.

In this work, we define a community to be a group of people who interact with resources (e.g. bookmarks) as well as with each other in a coherent manner. We model the social interaction among users as a *hypergraph* – where users are related to each other via one or more relations, and a relation involves two or more entities (e.g. bookmarking involves users, documents and tags). Our goal is to answer the following questions: (1) How to model multi-relational social data? (2) How to reveal the underlying communities consistent across multiple relations? (3) How to track those communities over time?

Related work. The structure of interactions among people have been modeled as clustering structure [2] in a bipartite graph. Multi-relational social network analysis concerns networks involving more than two types of entities. Existing techniques include tensor based analysis [2] or multi-graph mining [3], which do not take advantage of the relational sparsity in different context, or only deal with a particular relational context. We propose a flexible and efficient framework that exploits various relational context in social networks.

We introduce the *Relational Hypergraph Model* that captures the interactions between a set of facets (e.g. users, projects, etc.). We propose an analytic method, *Hypergraph Factorization*, to find consistent and soft communities on both static and time evolving

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

social data. Our method is based on novel nonnegative multi-tensor factorization with many desirable properties: (1) theoretically sound with convergence guarantee; (2) scalable to data size; (3) sparse solution that fits social network data; (4) probabilistic interpretation. We demonstrate the applicability of our method on real-world enterprise data, with excellent results.

2. APPROACH

Our approach consists of three main parts: (1) a relational hypergraph model, (2) a hypergraph factorization method, and (3) an on-line method to handle time-varying relational data.

Relational hypergraph model. Three concepts are involved in our model: *facet*, *relation*, and *relational hypergraph*. As illustrated in Figure 1(a), a set of users u_1, u_2, …, work closely with each other, under different working projects, j_1, j_2 ...; some of them share information (documents r_1, r_2, …, with a set of tags x_1, x_2, …) with others via bookmarking. To generally describe such collaboration data, we defines a *facet* as a set of objects or entities of the same type, e.g. a user facet is a set of users, and define a *relation* as the interactions among facets, e.g. (user, project) relation. A relation can involve two or more facets, e.g. “bookmark” (user, document, tag) is a 3-way relation.

We denote the q-th facet as $v^{(q)}$ and the set of all facets as V. An M-way relation e on facets $v^{(1)}$, $v^{(2)}$,…, $v^{(M)}$ is a subset of the Cartesian product $v^{(1)}\times\ldots\times v^{(M)}$. We denote a particular relation by $e^{(r)}$ where r is the relation index. The observations of an M-way relation $e^{(r)}$ is represented as an M-way data tensor $\boldsymbol{\mathcal{X}}^{(r)}$. We use a *relational hypergraph* to describe a combination of the relations of facets. A hypergraph is a graph where edges, called *hyperedges*, connect to any number of vertices. Figure 1(b) is a hypergraph corresponding to the data schema shown in Figure 1(a). For a set of facets V={$v^{(q)}$} and a set of relations E={$e^{(r)}$}, we can construct a *hypergraph* G=(V,E) where the vertices correspond to facets and hyperedges correspond to relations (ref. Figure 1(b), a hypergraph for the scenario (a)). To reduce notation complexity, V and E also represents the set of all vertex and edge indices respectively. A hyperedge/relation $e^{(r)}$ is said to be *incident* to a facet/vertex $v^{(q)}$ if $v^{(q)}\in e^{(r)}$, and can be represented by $v^{(q)}\sim e^{(r)}$ or $e^{(r)}\sim v^{(q)}$.

Hypergraph Factorization (HF). We seek to extract communities, i.e. groups of people who interact with each other in a coherent manner. The interactions are across multiple relations and multiple facets. For example, in Figure 1(c) we can use two communities to explain the interaction observed in Figure 1(a), as the interaction is mostly likely to occur within communities.

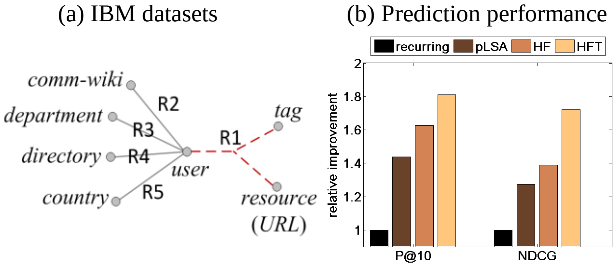
The key problem is how to extract the consistent clusters leveraging the hypergraph. We define the problem as *hypergraph factorization*: Given a hypergraph G=(V,E) and a set of data tensors $\{\boldsymbol{\mathcal{X}}^{(r)}\}_{r\in E}$ defined on G, find nonnegative core tensor [**z**] and factors $\{\mathbf{U}^{(q)}\}_{q\in V}$ for corresponding facets V={$v^{(q)}$}, where $\mathbf{U}^{(q)}$ is a $I_q\times K$ matrix, with its (i_q,k)-element indicating how likely an interaction in the k-th community involves the i-th user, and the elements of core tensor [**z**] indicate the prior probabilities of the communities. We solve the problem in terms of optimization – to approximate all data tensors by combining a common core tensor [**z**] and a common set of nonnegative factors $\{\mathbf{U}^{(q)}\}$. We have provided an efficient iterative algorithm that guarantees to find a (local) optimal solution. In our algorithm, the information contained in each relation is propagated to other relations via the core tensor and its connected facet factors.

Hypergraph Factorization with Time evolving data (HFT). When the relational data changes over time, we want the clustering structure to reflect these changes, but remain consistent with historic structure. We extend the problem to handle time evolving data through an incremental algorithm that updates community structure based on a prior community model.

3. EXPERIMENTAL RESULTS

We collected relational data from various social media used at IBM, as summarized in Figure 2(a). Note that relations can be static or dynamic, based on the update frequency of the data.

User interest prediction. In addition to qualitative case studies, we design a prediction task to illustrate how our community tracking algorithm can be utilized to predict users’ future interests. Specifically, given data D_t at time t, we predict users’ future use of tags, and compare the prediction with the truth data D_{t+1}. We use two information retrieval metrics, P@10 (the precision of the top 10 results) and NDCG (Normalized Discount Cumulative Gain) to compare our method with a frequency based method (denoted by “recuring”) and a collective filtering method (pLSA [1]). The results (ref. Figure 2(b)) indicate the prediction given by our approach outperforms the baseline methods by 36-250% on the average, which suggest that our method can better capture the dynamics of users’ interests. By leveraging cooperate relations, ~20% of the users whose future interests can be predicted.


(a) IBM datasets

(b) Prediction performance

Figure 2: (a) Hypergraph of IBM dataset: The gray solid edges represent static relations and the red dashed edges represent time-varying relation. The sizes of these relational data from R1~R5 are: 3K×12K×61K, 3K×1K, 3K×2K, 3K×90 and 3K×42. (b) Prediction performance: Our framework improves the prediction of users’ future tag use.

4. CONCLUSION

We introduce relational hypergraphs to model the schematic structures in social data in terms of facets and relations, and then solve the community extraction problem based on nonnegative multi-tensor decomposition on the hypergraph. The key contribution of our proposed method is its ability to flexibly handle both static and time evolving social contents with low time complexity. The experiment on real-world enterprise data suggests that our method captures the dynamics of users’ social context well.

5. REFERENCES

[1] T. Hofmann (1999). *Probabilistic latent semantic indexing*, SIGIR, 50-57,

[2] J. Sun, C. Faloutsos, S. Papadimitriou and P. Yu (2007). *GraphScope: parameter-free mining of large time-evolving graphs*, SIGKDD, 687-696, 2007.

[3] S. Zhu, K. Yu, Y. Chi and Y. Gong (2007). *Combining content and link for classification using matrix factorization*, SIGIR, 487-494, 2007.