Recommendation in Social Media
Recent Advance and New Frontiers

Jiliang Tang*, Jie Tang# and Huan Liu*
*Data Mining and Machine Learning Lab, Arizona State University
#Knowledge Engineering Lab, Tsinghua University

http://www.public.asu.edu/~jtang20/Recommendation.htm
August 24, 2014
Outline

- Introduction
- Friend Recommendation
- Content Recommendation
- Location Recommendation
- Summary
Outline

- Introduction
- Friend Recommendation
- Content Recommendation
- Location Recommendation
- Summary
Social Media [Zafarani et al., 2014]

- Social media greatly enables people to participate in online activities
  - Networking, tagging and commenting
- It shatters the barrier for online users to create and share information at any place at any time

Social media data increases at an unprecedented rate

- It becomes increasingly difficult for online users to get their interested information
Among 1,280,000,000 users, who should I befriend with in Facebook?

Facebook provides friend recommendations
– Suggesting ``People You May Know''
Who should I follow among 645,750,000 Twitter users?

Twitter implements a follow recommendation mechanism
– Suggesting ```Who to follow```
YouTube

100 hours of videos are uploaded into YouTube in every minute – How to find interested videos to watch?

In addition to searching with queries, YouTube also recommends some videos based on your browsing, searching and watching history.
Among hundreds of thousands of restaurants in New York City, which one I should go for dinner?

Yelp can suggest some restaurants based on their ratings and your current locations automatically.
Foursquare

- During a short visit in New York City, where should we go?

- Foursquare can suggest some places to visit and as some useful tips base on your locations automatically
Recommendation and Social Media

- Recommendation is widely used to mitigate information overload problem in social media.
- Social media and recommendation can mutually benefit each other [Guy and Carmel, 2011]

Social media introduces new types of data, advancing current recommendation research as well as expanding research frontiers.

Recommendation suggests social media users relevant information, significantly impacting the success of social media.
Recommendation in Social Media

- Social media users can be described with three types of information
  - Social information
  - Content information
  - Location information

- Three information types correspond to three recommendation tasks
  - Friend Recommendation
  - Content Recommendation
  - Location Recommendation
Special Characteristics of Recommendation

- Search starts with a user’s explicit query
- Recommendation is triggered with a user’s implicit query
- A user is provided with relevant and timely information without explicitly stating his needs
- Why is this needed and critical?
  - If successful, a user will choose to stay longer on the site, or pay more attention, leading to purchasing more products, contributing more
Recommendation in Social Media

- Friend Recommendation in Social Media
- Content Recommendation in Social Media
- Location Recommendation in Social Media
- Performance Evaluations
Recommendation in Social Media

- Friend Recommendation in Social Media
- Content Recommendation in Social Media
- Location Recommendation in Social Media
- Performance Evaluations
Social information is usually represented as user–user matrix $S$

Friend recommendation is basically to predict missing links

– Supervised and unsupervised methods
Supervised Approaches [Lichtenwalter et al., 2010]

- Supervised approaches consider link prediction as a classification problem

- Label preparation
  - Existence of links as labels

- Feature extraction
  - Extracting a set of features from available sources to represent pairs of users
Unsupervised Approaches

- Unsupervised methods are usually based on the characteristics of the given networks

- Connectivity based methods [David and Jon et al., 2007]
  - Common neighbors
  - Jaccard’s coefficient
  - Adamic/Adar

- Low-rank matrix factorization based methods
  [Menon and Elkan, 2011]
Friend Recommendation in Social Media

- The availability of large social networks in social media allows recent efforts on understanding the mechanism of the dynamic formation of friendships

- Reciprocity recommendation
  - A follows B, how likely B will follow A back?

- Triadic closure recommendation
  - What are the fundamental factors that trigger the formation of triadic closure?

- Cross-community recommendation
  - Analyzing the cross-community friendship formation problem and uncovering the underlying patterns
Recommenda)on in Social Media

Friend Recommenda)on in Social Media

Content Recommenda)on in Social Media

Loca)on Recommenda)on in Social Media

Performance Evaluations
Fundamental Recommendation Approaches
[Adomavicius and Tuzhilin, 2005]

- User and content item relation can be represented as an user-item matrix \( R \)

  ![User and Item Diagram]

- Content-based recommendation
  - Recommending items similar to the ones that the user has preferred in the past

- Collaborative filtering (CF)–based recommendation
  - Using the user's past behavior to uncover user preferences
  - Memory-based CF and Model-based CF

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
Memory-based Collaborative Filtering

- It uses either the whole user-item matrix or a sample to generate a prediction
  - Needing memory to store the user-item matrix
- User-oriented collaborative filtering
  - Calculating user-user similarity
  - Aggregating ratings from similar users
- Item-oriented collaborative filtering
  - Computing item-item similarity
  - Aggregating ratings from similar items
User-oriented collaborative Filtering

- Calculating user-user similarity
  - Cosine similarity

\[ S(i, j) = \frac{\sum_{k \in I} R_{ik} R_{jk}}{\sqrt{\sum_{k \in I} R_{ik}^2} \sqrt{\sum_{k \in I} R_{jk}^2}} \]

- Aggregating ratings from similar users

\[ \hat{R}_{ij} = \frac{\sum_{u_k \in N_i} S_{ik} R_{kj}}{\sum_{u_k \in N_i} S_{ik}} \]

- $I$ is the set of items rated by $u_i$ and $u_j$
- $R_{ik}$ is the rating to the $k$th item from $u_i$
- $N_i$ is the set of users who have rated the $j$-th item
An Illustration of User-oriented Collaborative Filtering

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

1, 2, and 3 are users
A, B, C, D, and E are items
R(1,D) = ?

- **Calculating cosine similarity**
  
  \[
  S(1,2) = \frac{3 \times 3 + 4 \times 4}{\sqrt{3 \times 3 + 4 \times 4} \sqrt{3 \times 3 + 4 \times 4}} = 1
  \]
  
  \[
  S(1,3) = \frac{5 \times 1 + 4 \times 2}{\sqrt{5 \times 5 + 4 \times 4} \sqrt{1 \times 1 + 2 \times 2}} = 0.9080
  \]

- **Aggregating ratings**

  \[
  \hat{R}(1,D) = \frac{R(2,D) \times S(1,2) + R(3,D) \times S(1,3)}{S(1,2) + S(1,3)}
  \]
  
  \[
  = \frac{4 \times 1 + 2 \times 0.9080}{1 + 0.9080} = 3.05
  \]
**Model-based Collaborative Filtering**

- It assumes there exists a model that generates the ratings and the model parameters can be learned
  - Storing only parameters instead of the rating matrix
  - Using the assumed model with parameters to do prediction

- Matrix factorization methods are very competitive and are widely adopted to build recommender systems [Koren et al., 2009]

\[
R_{ij} = U_i V_j^T
\]
An Illustration of Matrix Factorization based CF

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>?</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

1, 2 and 3 are users
A, B, C, D, and E are items
The latent dimension $k = 1$
$R(1,D) = ?$

- Learning Latent Factors $U$ and $V$
  
  $U = \begin{pmatrix} 2.4109 \\ 1.4706 \end{pmatrix}$
  
  $V = \begin{pmatrix} 1.5740 \\ 1.5990 \\ 2.5716 \end{pmatrix}$

- Reconstructing the rating matrix

\[ \hat{R} = UV^T = \begin{pmatrix} 4.2756 & 3.2047 & 4.1408 & 4.2066 & 6.7654 \\ 3.9182 & 2.9368 & 3.7946 & 3.8550 & 6.1999 \\ 2.3901 & 1.7915 & 2.3147 & 2.3515 & 3.7819 \end{pmatrix} \]
Content Recommendation in Social Media

- New types of data introduced by social media have greatly enriched the sources available for content recommendation
  - Social information
  - Location information

- Content recommendation with social networks
  - How to put social information into content recommendation?

- Location-aware content recommendation
  - Given the locations of users, how to recommend their interested content?
A number of location-based social networking services have emerged in recent years — Foursquare, Yelp, and Facebook Places.

Location recommendation is to recommend to a user some POIs for his future visits based on his LBSN context.
Recommendation in Social Media

- Friend Recommendation in Social Media
- Content Recommendation in Social Media
- Location Recommendation in Social Media
- Performance Evaluations
Recommendation Evaluations

- Different evaluation metrics assess recommender systems from different perspectives
  - Prediction power: the ability to accurately predict users’ choices
  - Classification accuracy: the ability to differentiate good items from bad ones
  - Novelty and exploration: the ability to discover new items or explore diverse items
Prediction Accuracy Evaluation

- Prediction Accuracy Evaluation measures the average error of predicted ratings
  - Mean Absolution Error (MAE)
  \[ MAE = \frac{\sum_{<u_i, v_j> \in O} |\hat{R}_{ij} - R_{ij}|}{|O|} \]
  - Root Mean Squared Error (RMSE)
  \[ RMSE = \sqrt{\frac{\sum_{<u_i, v_j> \in O} (\hat{R}_{ij} - R_{ij})^2}{|O|}} \]
Ranking Accuracy Evaluation

- Ranking Accuracy evaluates how many recommended items are acquired by the users.

- **Precision@N**
  - How many top-N recommended items are acquired?
  - For a target user $u_i$
  
  \[
  \text{Precision@N} = \frac{|\text{TopN}(i) \cap L(i)|}{|\text{TopN}(i)|}
  \]

- **Recall@N**
  - How many top-N acquired items are recommended?
  - For a target user $u_i$
  
  \[
  \text{Recall@N} = \frac{|\text{TopN}(i) \cap L(i)|}{|L(i)|}
  \]
Coverage Evaluation

- **Item coverage**
  - Evaluating how good the items recommended by a recommendation system $S$ are
  
  \[ I_c = \frac{|N_d|}{|N|} \]

  $N$ is the set of items supposed to be recommended, while $N_d$ is the set of items recommended by $S$

- **User coverage**
  - Evaluating how good the users recommended by a recommendation system $S$ are
  
  \[ U_c = \frac{|M_d|}{|M|} \]

  $M$ is the set of users supposed to be recommended, while $M_d$ is the set of users $S$ recommends
References


Outline

Introduction
Friend Recommendation
Content Recommendation
Location Recommendation
Summary
Networked World

- **1.3 billion** users
- **700 billion** minutes/month

- **600 million** users
- **.5 billion** tweets/day

- **79 million** users per month
- **9.65 billion** items/year

- **280 million** users
- **80% of users** are 80-90’s

- **560 million** users
- **influencing** our daily life

- **500 million** users
- **35 billion** on 11/11

- **800 million** users
- **~50% revenue** from network life
Social Tie Recommendation

Inferring social ties

Reciprocity

Triadic Closure

KDD 2010, PKDD 2011 (Best Paper Runnerup), WSDM 2012, ACM TKDD
Inferring Social Ties

KDD 2010, PKDD 2011 (*Best Paper Runner-up*), WSDM 2012, DMKD
Real social networks are complex...

- Nobody exists only in one social network.
  - Public network vs. private network
  - Business network vs. family network
- However, existing networks (e.g., Facebook and Twitter) are trying to lump everyone into one big network
  - FB tries to solve this problem via lists/groups
  - However...
- Google+

which circle? Users do not take time to create it.
Even complex than we imaged!

- Only 16% of mobile phone users in Europe have created custom contact groups
  - *users do not* take the time to create it
  - *users do not* know how to circle their friends

- The fact is that our social network is **black-white...**
Example 1: finding **boss** in email networks

(PKDD’11, Best Paper Runnerup)

User interactions may form *implicit groups*

Enterprise email network

How to infer:

- CEO
- Manager
- Employee
Example 2: finding friends in mobile networks

From Home 08:40
From Office 11:35
From Office 15:20
From Office 17:55
From Outside 21:30
Both in office 08:00 – 18:00

Friends
Other

0.89 0.98
0.77 0.63
0.86 0.70

0.77 0.63
0.86 0.70
Challenges

- What are the **fundamental forces** behind?
- Can we automatically infer the type of social ties?

Publication network

Twitter’s following network

Mobile communication network
Networks

- **Epinions** a network of product reviewers: 131,828 nodes (users) and 841,372 edges
  - trust relationships between users

- **Slashdot**: 82,144 users and 59,202 edges
  - “friend” relationships between users

- **Mobile**: 107 mobile users and 5,436 edges
  - to infer friendships between users

- **Coauthor**: 815,946 authors and 2,792,833 coauthor relationships
  - to infer advisor-advisee relationships between coauthors

- **Enron**: 151 Enron employees and 3572 edges
  - to infer manager-subordinate relationships between users.
Problem Formulation

Input: $G=(V, E^L, E^U, R^L, W)$

- $V$: Set of Users
- $E^L, R^L$: Labeled relationships
- $E^U$: Unlabeled relationships
- $W$: Attributes

Output: $f: G \rightarrow R$
Basic Idea

Recommendation in Social Media

Arizona State University
Tsinghua University

KDD2014 Tutorial 47
Partially Labeled Pairwise Factor Graph Model (PLP-FGM)

**Constraint factor** $h$

**Latent Variable**

**Correlation factor** $g$

**Attribute factors** $f$

**Problem:**
For each relationship, identify which type has the highest probability?

**Example:**
A makes call to B immediately after the call to C.

Different ways to instantiate factors

- We use exponential-linear functions
  
  \[ f(y_i, x_i) = \frac{1}{Z_{\lambda}} \exp\{\lambda^T \Phi(y_i, x_i)\} \]

- Attribute Factor:

- Correlation / Constraint Factor:

\[ g(y_i, G(y_i)) = \frac{1}{Z_\alpha} \exp\{ \sum_{y_j \in G(y_i)} \alpha^T g(y_i, y_j) \} \]
\[ h(y_i, H(y_i)) = \frac{1}{Z_\beta} \exp\{ \sum_{y_j \in H(y_i)} \beta^T h(y_i, y_j) \} \]

- \( \theta = [\lambda, \alpha, \beta], s = [\Phi^T, g^T, h^T]^T \)
- Log-Likelihood of labeled Data:

\[ \mathcal{O}(\theta) = \log \sum_{Y|Y^L} \exp\{\theta^T S\} - \log \sum_Y \exp\{\theta^T S\} \]
Maximize the log-likelihood of labeled relationships

**Algorithm 1: Learning PLP-FGM.**

**Input:** learning rate $\eta$

**Output:** learned parameters $\theta$

Initialize $\theta$;

repeat

- Calculate $E_{P_{\theta}(Y|Y^L,G)}S$ using LBP;
- Calculate $E_{P_{\theta}(Y|G)}S$ using LBP;
- Calculate the gradient of $\theta$ according to Eq. 7:

\[
\nabla \theta = E_{P_{\theta}(Y|Y^L,G)}S - E_{P_{\theta}(Y|G)}S
\]

Update parameter $\theta$ with the learning rate $\eta$:

\[
\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \nabla \theta
\]

until Convergence;

Gradient Ascent Method

Code: [http://arnetminer.org/socialtie](http://arnetminer.org/socialtie)
Questions:
- How to obtain sufficiently training data?
- Can we leverage knowledge from other network?
Social Theories

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

Observations:
(1) The underlying networks are unbalanced;
(2) While the friendship networks are balanced.
Social Theories—Structural hole

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

**Observations:** Users are more likely (+25-150% higher than change) to have the same type of relationship with C if C spans structural holes.
Social Theories—Social status

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

Observations: 99% of triads in the networks satisfy the social status theory

Note: Given a triad (A,B,C), let us use 1 to denote the advisor-advisee relationship and 0 colleague relationship. Thus the number 011 to denote A and B are colleagues, B is C’s advisor and A is C’s advisor.
Social Theories—Two-step-flow

- Social balance theory
- Structural hole theory
- Social status theory
- Two-step-flow theory

**Observations:** Opinion leaders are more likely (+71%-84% higher than chance) to have a higher social-status than ordinary users.

**OL**: Opinion leader;
**OU**: Ordinary user.
Transfer Factor Graph Model

Coauthor network

Output: social network

Observations

Bridge via social theories

Mobile

Observations

TrFG model

Coauthor network

Mobile

TrFG model
Mathematical Formulation

\[
\mathcal{O}(\alpha, \beta, \mu) = \mathcal{O}_S(\alpha, \mu) + \mathcal{O}_T(\beta, \mu)
\]

\[
= \sum_{i=1}^{|V_S|} \sum_{j=1}^{d} \alpha_{ij} g_j(x_{ij}^S, y_{ij}^S) + \sum_{i=1}^{d'} \sum_{j=1}^{d'} \beta_{ij} g'_j(x_{ij}^T, y_{ij}^T)
\]

\[
+ \sum_{k} \mu_k \left( \sum_{c \in G_S} h_k(Y_c^S) + \sum_{c \in G_T} h_k(Y_c^T) \right)
\]

\[- \log Z \]

Features defined in source network
Features defined in target network
Triad-based features shared across networks

Experiments

- **Data sets**
  - **Epinions**: 131,828 nodes (users) and 841,372 edges
  - **Slashdot**: 82,144 users and 59,202 edges
  - **Mobile**: 107 mobile users and 5,436 edges
  - **Coauthor**: 815,946 authors and 2,792,833 coauthor relationships
  - **Enron**: 151 Enron employees and 3572 edges

- **Comparison methods**
  - **SVM** and **CRF** are two baseline methods
  - **PFG** is the partially-labeled factor graph model
  - **TranFG** is the transfer–based factor graph model
### Results – undirected networks

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epinions (S) to Slashdot (T) (40%)</td>
<td>SVM</td>
<td>0.7157</td>
<td>0.9733</td>
<td>0.8249</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.8919</td>
<td>0.6710</td>
<td>0.7658</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>0.9300</td>
<td>0.6436</td>
<td>0.7607</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.9414</td>
<td>0.9446</td>
<td>0.9430</td>
</tr>
<tr>
<td>Slashdot (S) to Epinions (T) (40%)</td>
<td>SVM</td>
<td>0.9132</td>
<td>0.9925</td>
<td>0.9512</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.8923</td>
<td>0.9911</td>
<td>0.9393</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>0.9954</td>
<td>0.9787</td>
<td>0.9870</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.9954</td>
<td>0.9787</td>
<td>0.9870</td>
</tr>
<tr>
<td>Epinions (S) to Mobile (T) (40%)</td>
<td>SVM</td>
<td>0.8983</td>
<td>0.5955</td>
<td>0.7162</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.9455</td>
<td>0.5417</td>
<td>0.6887</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>1.0000</td>
<td>0.5924</td>
<td>0.7440</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.8239</td>
<td>0.8344</td>
<td>0.8291</td>
</tr>
<tr>
<td>Slashdot (S) to Mobile (T) (40%)</td>
<td>SVM</td>
<td>0.8983</td>
<td>0.5955</td>
<td>0.7162</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.9455</td>
<td>0.5417</td>
<td>0.6887</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>1.0000</td>
<td>0.5924</td>
<td>0.7440</td>
</tr>
</tbody>
</table>
|                           | TranFG | 0.7258 | 0.8599 | 0.7872   

**SVM and CRF** are two baseline methods.

**PFG** is the proposed partially-labeled factor graph model.

**TranFG** is the proposed transfer-based factor graph model.
Results – directed networks

**SVM** and **CRF** are two baseline methods.

**PFG** is the proposed partially-labeled factor graph model.

**TranFG** is the proposed transfer–based factor graph model.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Prec.</th>
<th>Rec.</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coauthor (S) to Enron (T) (40%)</td>
<td>SVM</td>
<td>0.9524</td>
<td>0.5556</td>
<td>0.7018</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>0.9565</td>
<td>0.5366</td>
<td>0.6875</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td><strong>0.9730</strong></td>
<td>0.6545</td>
<td>0.7826</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.9556</td>
<td><strong>0.7818</strong></td>
<td><strong>0.8600</strong></td>
</tr>
<tr>
<td>Enron (S) to Coauthor (T) (40%)</td>
<td>SVM</td>
<td>0.6910</td>
<td>0.3727</td>
<td>0.4842</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td><strong>1.0000</strong></td>
<td>0.3043</td>
<td>0.4666</td>
</tr>
<tr>
<td></td>
<td>PFG</td>
<td>0.9916</td>
<td>0.4591</td>
<td>0.6277</td>
</tr>
<tr>
<td></td>
<td>TPFG</td>
<td>0.5936</td>
<td><strong>0.7611</strong></td>
<td>0.6669</td>
</tr>
<tr>
<td></td>
<td>TranFG</td>
<td>0.9793</td>
<td>0.5525</td>
<td><strong>0.7065</strong></td>
</tr>
</tbody>
</table>
Social Tie Recommendation

Inferring social ties

Reciprocity

You ➔ Lady Gaga ➔ You

Triadic Closure

You ➔ Lady Gaga ➔ You

Lady Gaga ➔ Shiteng ➔ Lady Gaga

KDD 2010, PKDD 2011 (Best Paper Runnerup), WSDM 2012, ACM TKDD
Parasocial vs. Reciprocal

Who will follow you back?

On Twitter...

- Ladygaga (100%)
- Obama (100%)
- Shiteng (1%)
- Huwei (60%)
- JimmyQiao (30%)
Homophily

**Link homophily:** Users who share common links will have a tendency to follow each other.

**Status homophily:** Elite users have a much stronger tendency to follow each other.
Retweet vs. reply

*Retweeting seems to be more helpful
Structural Balance

(A) and (B) are balanced, but (C) and (D) are not.

- Structural balance
  - Reciprocal relationships are balanced (88%);
  - Parasocial relationships are not (only 29%).
Triad Factor Graph (TriFG)

Input: Mobile Network

Observations

TriFG model

$y_1 = \text{friend}$

$y_2 = \text{friend}$

$y_4 = \text{non-friend}$

$y_5 = \text{non-friend}$

$h(y_3, y_4, y_5)$

$f(v_4^u, v_1^s, y_1)$

$f(v_2^u, v_2^s, y_2)$

$f(v_3^u, v_3^s, y_3)$

$f(v_4^u, v_4^s, y_4)$

$f(v_5^u, v_5^s, y_5)$

$f(v_6^u, v_6^s, y_6)$

$(v_2, v_3)$

$(v_4, v_3)$

$(v_6, v_5)$

$(v_4, v_6)$

$(v_2, v_1)$

$(v_3, v_1)$

$(v_5, v_1)$

$(v_6, v_1)$
Experiments

- Huge sub-network of twitter
  - 13,442,659 users and 56,893,234 following links.
  - Extracted 35,746,366 tweets.

- Dynamic networks
  - With an average of 728,509 new links per day.
  - Averagely 3,337 new follow-back links per day.
  - 13 time stamps by viewing every four days as a time stamp.

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1Measure</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Case 1</td>
<td>SVM</td>
<td>0.6908</td>
<td>0.6129</td>
<td>0.6495</td>
<td>0.9590</td>
</tr>
<tr>
<td></td>
<td>LRC</td>
<td>0.6957</td>
<td>0.2581</td>
<td>0.3765</td>
<td>0.9510</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>1.0000</td>
<td>0.6290</td>
<td>0.7723</td>
<td>0.9770</td>
</tr>
<tr>
<td></td>
<td>TriFG</td>
<td>1.0000</td>
<td>0.8548</td>
<td><strong>0.9217</strong></td>
<td><strong>0.9910</strong></td>
</tr>
<tr>
<td>Test Case 2</td>
<td>SVM</td>
<td>0.7323</td>
<td>0.6212</td>
<td>0.6722</td>
<td>0.9534</td>
</tr>
<tr>
<td></td>
<td>LRC</td>
<td>0.8333</td>
<td>0.3030</td>
<td>0.4444</td>
<td>0.9417</td>
</tr>
<tr>
<td></td>
<td>CRF</td>
<td>1.0000</td>
<td>0.6333</td>
<td>0.7755</td>
<td>0.9717</td>
</tr>
<tr>
<td></td>
<td>TriFG</td>
<td>1.0000</td>
<td>0.8788</td>
<td><strong>0.9355</strong></td>
<td><strong>0.9907</strong></td>
</tr>
</tbody>
</table>
Effect of Time Span

- Distribution of follow back time
  - 60% for next-time stamp;
  - 37% for following 3 time stamps.

- Different settings of the time span
  - Performance drops sharply when two or less;
  - Acceptable for three time stamps.
Case Study

(a) Ground Truth

(b) SVM

(c) Our approach (TriFG)
Social Tie Recommendation

Inferring social ties

Reciprocity

Triadic Closure

Triadic Closure

Tiancheng Lou, Jie Tang, John Hopcroft, Zhanpeng Fang, Xiaowen Ding. Learning to Predict Reciprocity and Triadic Closure in Social Networks. ACM TKDD, V(7) 2, July 2013, Article No. 5.
Triadic Closure

Ladygaga <-> Obama: 0.5%
Ladygaga <-> Shiteng: 0.6%
Ladygaga <-> Huwei: 1%
Shiteng <-> Obama: 1%
Shiteng <-> Huwei: 60%
Huwei <-> Obama: 50%

JimmyQiao
Triads in networks

Open Triad to Triadic Closure

Hong Huang, Jie Tang, Sen Wu, Lu Liu, and Xiaoming Fu. Mining Triadic Closure Patterns in Social Networks. In WWW'14, pages 499-504.
Weibo Data


- 700,000 nodes
- 400M following links
- 360,000 new links
- 44,000 newly formed closed triads per day
- 200 out-degree per user

Weibo Data
Open Triad to Triadic Closure

Y-axis: probability that each open triad forms triadic closures
Observations: 1. Women attract people to close triads;
2. Men are more inclined to form triadic closure;
3. Men are important to bridge people.
Social Role

0—ordinary user
1—opinion leader (top 1% PageRank)
e.g., 001 means A and B are ordinary user while C is opinion leader.

Observations: Triads of opinion leaders themselves are more likely to be closed.
Triad Factor Graph (TriFG)

Input: Mobile Network

TriFG model

Observations

\[ y_1 = \text{friend} \]

\[ y_2 = \text{friend} \]

\[ y_3 = \text{?} \]

\[ y_4 = \text{?} \]

\[ y_5 = \text{non-friend} \]

\[ y_6 = \text{non-friend} \]

\[ \text{Inpout: Mobile Network} \]

\[ f(v_1^u, v_1^s, y_1) \]

\[ f(v_2^u, v_1^s, y_2) \]

\[ f(v_3^u, v_3^s, y_3) \]

\[ f(v_4^u, v_4^s, y_4) \]

\[ f(v_5^u, v_5^s, y_5) \]

\[ f(v_6^u, v_6^s, y_6) \]

\[ h(y_1, y_2, y_3) \]

\[ h(y_3, y_4, y_5) \]
## Triad Closure Prediction Result

<table>
<thead>
<tr>
<th>Data</th>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Test Case 1</strong></td>
<td><strong>SVM</strong></td>
<td>0.0870</td>
<td>0.1429</td>
<td>0.1081</td>
</tr>
<tr>
<td></td>
<td><strong>LRC</strong></td>
<td>0.0536</td>
<td>0.1404</td>
<td>0.0759</td>
</tr>
<tr>
<td></td>
<td><strong>CRF-balance</strong></td>
<td>0.0208</td>
<td>0.0436</td>
<td>0.0282</td>
</tr>
<tr>
<td></td>
<td><strong>CRF</strong></td>
<td>0.1111</td>
<td>0.0870</td>
<td>0.0976</td>
</tr>
<tr>
<td></td>
<td><strong>wTriFG</strong></td>
<td>0.3333</td>
<td>0.0373</td>
<td>0.0671</td>
</tr>
<tr>
<td></td>
<td><strong>TriFG</strong></td>
<td><strong>0.4545</strong></td>
<td><strong>0.2174</strong></td>
<td><strong>0.2941</strong></td>
</tr>
<tr>
<td><strong>Test Case 2</strong></td>
<td><strong>SVM</strong></td>
<td>0.2000</td>
<td>0.2222</td>
<td>0.2105</td>
</tr>
<tr>
<td></td>
<td><strong>LRC</strong></td>
<td>0.1071</td>
<td>0.1667</td>
<td>0.1304</td>
</tr>
<tr>
<td></td>
<td><strong>CRF-balance</strong></td>
<td>0.0909</td>
<td>0.0556</td>
<td>0.0690</td>
</tr>
<tr>
<td></td>
<td><strong>CRF</strong></td>
<td>0.2222</td>
<td>0.2222</td>
<td>0.2222</td>
</tr>
<tr>
<td></td>
<td><strong>wTriFG</strong></td>
<td>0.5000</td>
<td>0.0556</td>
<td>0.1000</td>
</tr>
<tr>
<td></td>
<td><strong>TriFG</strong></td>
<td><strong>0.8571</strong></td>
<td><strong>0.3333</strong></td>
<td><strong>0.4800</strong></td>
</tr>
</tbody>
</table>
Cross-domain Collaboration

Interdisciplinary collaborations have generated huge impact, for example,

- 51 (>1/3) of the KDD 2012 papers are result of cross-domain collaborations between graph theory, visualization, economics, medical inf., DB, NLP, IR
- Research field evolution
Increasing trend of cross-domain collaborations

Data Mining (DM), Medical Informatics (MI), Theory (TH), Visualization (VIS)
Challenges

1. Sparse Connection: <1%
2. Complementary expertise
3. Topic skewness: 9%

Data Mining
- Sparse Connection: <1%
- Large graph
- Heterogeneous network
- Social network

Theory
- Graph theory
- Automata theory
- Complexity theory

Complementary expertise:
- Large graph
- Heterogeneous network
- Social network

Recommendation in Social Media

KDD2014 Tutorial
Author Matching

Data Mining

Medical Informatics

Author

Cross-domain coauthorships

Query user

Sparse connection!

\[ r^{(t+1)} = (1 - \tau)S \cdot r^{(t)} + \tau q \]
**Topic Matching**

Data Mining

\[ G^S \]

\[ v_1 \]
\[ v_2 \]
\[ \ldots \]
\[ v_N \]
\[ v_q \]

Topics Extraction

\[ Z_1 \]
\[ Z_2 \]
\[ Z_3 \]
\[ \ldots \]
\[ Z_T \]

\[ Z'_1 \]
\[ Z'_2 \]
\[ Z'_3 \]
\[ \ldots \]
\[ Z'_T \]

Medical Informatics

Topics correlations

2. Complementary Expertise!
3. Topic skewness!
Identify “cross-domain” Topics

Data Mining

$G^S$

$v_1$

$v_2$

$\ldots$

$v_N$

$v_q$

Topics

$Z_1$

$Z_2$

$Z_3$

$\ldots$

$Z_K$

Medical Informatics

$G^T$

$v'_1$

$v'_2$

$\ldots$

$v'_{N'}$
Collaboration Topics Extraction

Step 1:
- Initialize an ACT model in \( G^S \) by learning from documents written by authors only from \( G^S \).
- Similarly, initialize an ACT model for target domain \( G^T \).

Step 2:
- For each collaboratively written document \( d \):
  - For each word \( x_{di} \) in \( d \):
    - Toss a coin \( s_{di} \) according to \( \text{bernoulli}(s_{di}) \sim \text{beta}(\gamma_t, \gamma) \), where \( \text{beta}(.) \) is a Beta distribution, and \( \gamma_t \) and \( \gamma \) are two parameters.
    - If \( s_{di} = 0 \) then:
      - Randomly select a pair \( (v, v') \) from \( d \)'s authors, where \( v \) is an author from \( G^S \) and \( v' \) from \( G^T \).
      - Draw a topic \( z_{di} \sim \text{multi}(\theta_{vv'}) \) from the topic mixture \( \theta_{vv'} \) specific to \( (v, v') \).
    - If \( s_{di} = 1 \) then:
      - Randomly select a user \( v \).
      - Draw a topic \( z_{di} \sim \text{multi}(\theta_{v}) \) from the topic model of user \( v \).
  - Draw a word \( x_{di} \sim \text{multi}(\phi z_{di}) \) from \( z_{di} \)-specific word distribution.
Intuitive explanation of Step 2 in CTL
Data Set and Baselines

• Arnetminer (available at http://arnetminer.org/collaboration)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Authors</th>
<th>Relationships</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining</td>
<td>6,282</td>
<td>22,862</td>
<td>KDD, SDM, ICDM, WSDM, PKDD</td>
</tr>
<tr>
<td>Medical Informatics</td>
<td>9,150</td>
<td>31,851</td>
<td>JAMIA, JBI, AIM, TMI, TITB</td>
</tr>
<tr>
<td>Theory</td>
<td>5,449</td>
<td>27,712</td>
<td>STOC, FOCS, SODA</td>
</tr>
<tr>
<td>Visualization</td>
<td>5,268</td>
<td>19,261</td>
<td>CVPR, ICCV, VAST, TVCG, IV</td>
</tr>
<tr>
<td>Database</td>
<td>7,590</td>
<td>37,592</td>
<td>SIGMOD, VLDB, ICDE</td>
</tr>
</tbody>
</table>

• Baselines
  – Content Similarity(Content)
  – Collaborative Filtering(CF)
  – Hybrid
  – Katz
  – Author Matching(Author), Topic Matching(Topic)
# Performance Analysis

## Training: collaboration before 2001  
### Validation: 2001-2005

<table>
<thead>
<tr>
<th>Cross Domain</th>
<th>ALG</th>
<th>P@10</th>
<th>P@20</th>
<th>MAP</th>
<th>R@100</th>
<th>ARHR -10</th>
<th>ARHR -20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Mining(S) to Theory(T)</td>
<td>Content</td>
<td>10.3</td>
<td>10.2</td>
<td>10.9</td>
<td>31.4</td>
<td>4.9</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>CF</td>
<td>15.6</td>
<td>13.3</td>
<td>23.1</td>
<td>26.2</td>
<td>4.9</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>Hybrid</td>
<td>17.4</td>
<td>19.1</td>
<td>20.0</td>
<td>29.5</td>
<td>5.0</td>
<td>2.4</td>
</tr>
<tr>
<td></td>
<td>Author</td>
<td>27.2</td>
<td>22.3</td>
<td>25.7</td>
<td>32.4</td>
<td>10.1</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td>Topic</td>
<td>28.0</td>
<td>26.0</td>
<td>32.4</td>
<td>33.5</td>
<td>13.4</td>
<td>7.1</td>
</tr>
<tr>
<td></td>
<td>Katz</td>
<td>30.4</td>
<td>29.8</td>
<td>21.6</td>
<td>27.4</td>
<td>11.2</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>CTL</td>
<td>37.7</td>
<td>36.4</td>
<td>40.6</td>
<td>35.6</td>
<td>14.3</td>
<td>7.5</td>
</tr>
</tbody>
</table>

**Content Similarity (Content):** based on similarity between authors’ publications  
**Collaborative Filtering (CF):** based on existing collaborations  
**Hybrid:** a linear combination of the scores obtained by the Content and the CF methods.  
**Katz:** the best link predictor in link-prediction problem for social networks  
**Author Matching (Author):** based on the random walk with restart on the collaboration graph  
**Topic Matching (Topic):** combining the extracted topics into the random walking algorithm
CTL can still maintain about 0.3 in terms of MAP which is significantly higher than baselines.
Parameter Analysis

(a) varying the number of topics $T$
(b) varying $\alpha$ parameter
(c) varying the restart parameter $\tau$ in the random walk
(d) Convergence analysis
Prototype System
http://arnetminer.org/collaborator

Cross-Domain Collaboration Recommendation

Treemap: representing subtopic in the target domain

Recommend Collaborators & Their relevant publications
Application in Game Networks
Applying Social Tie in Game Data

- Online gaming is one of the largest industries on the Internet...

- Facebook
  - 250 million users play games monthly
  - 200 games with more than 1 million active users
  - 12% of the company’s revenue is from games

- Tencent (Market Cap: ~150B $)
  - More than 400 million gaming users
  - 50% of Tencent’s overall revenue is from games

Zhanpeng Fang, Xinyu Zhou, Jie Tang, Wei Shao, A.C.M. Fong, Longjun Sun, Ying Ding, Ling Zhou, and Jarder Luo. Modeling Paying Behavior in Online Social Networks. In CIKM'14.
Two games: DNF

- Dungeon & Fighter Online (DNF)
  - A game of melee combat between users and large number of underpowered enemies
  - 400+ million users, the 2\textsuperscript{nd} largest online game in China
  - Users in the game can fight against enemies by individuals or by groups
Two games: QQ Speed

- QQ Speed
  - A racing game that users can partake in competitions to play against other users
  - 200+ million users
  - Users can race against other users by individuals or forma a group to race together
  - Some users may pay...
Task

• Given behavior log data and paying logs of online game users, predict

  Free users -> Paying users

• Will social influence play an important role in this task?
## The Big social data

- **Statistics of the datasets**

<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>QQSpeed</th>
<th>DNF</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>free users</td>
<td>5,812,894</td>
<td>204,112</td>
</tr>
<tr>
<td></td>
<td>paying users</td>
<td>1,394,630</td>
<td>109,099</td>
</tr>
<tr>
<td></td>
<td>new payers</td>
<td>399,747</td>
<td>34,568</td>
</tr>
<tr>
<td>Relationship</td>
<td>co-playing</td>
<td>134,812,639</td>
<td>7,306,265</td>
</tr>
<tr>
<td>Guild</td>
<td>guilds</td>
<td>600,860</td>
<td>49,680</td>
</tr>
<tr>
<td></td>
<td>co-guild</td>
<td>66,740,051</td>
<td>51,792,212</td>
</tr>
<tr>
<td>Activity</td>
<td>activity types</td>
<td>58</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>activity logs</td>
<td>44,742,907,507</td>
<td>5,716,434,808</td>
</tr>
<tr>
<td>Date span</td>
<td>from</td>
<td>2013.6.20</td>
<td>2013.4.1</td>
</tr>
<tr>
<td></td>
<td>to</td>
<td>2013.8.20</td>
<td>2013.6.30</td>
</tr>
</tbody>
</table>
Demographics Analysis

![Graph showing hazard ratio vs. gender and level distribution]
Analysis – Social influence

• Social network construction
  – Co-playing network

• Social relationship
  – Social influence
  – Strong/Weak tie
  – Status

• Structural influence
Social Influence

![Graph showing the probability of influence based on the number of paying neighbors. The graph illustrates the increase in probability with the number of neighbors, with significant jumps at certain thresholds (x3 and x5).]
Influence + Tie Strength

Graph showing the probability of influence as a function of the number of paying neighbors, with different tie strengths indicated by different colors.
Structural Influence

The prediction of feature vector $\mathbf{x}_i$:

$$y(\mathbf{x}_i) = w_0 + \sum_{j=1}^{d} w_j x_{ij} + \sum_{j=1}^{d-1} \sum_{j'=j+1}^{d} x_{ij} x_{ij'} \langle \mathbf{p}_j, \mathbf{p}_{j'} \rangle$$

and the corresponding objective function:

$$O = \sum_{x_i} (y(\mathbf{x}_i) - y_i)^2 + \lambda \sum_{i=1}^{d} \| \mathbf{p}_i \|^2$$

Zhanpeng Fang, Xinyu Zhou, Jie Tang, Wei Shao, A.C.M. Fong, Longjun Sun, Ying Ding, Ling Zhou, and Jarder Luo. Modeling Paying Behavior in Online Social Networks. In CIKM'14.
Online Test

- Test setting
  - Two groups: **test group** and **control group**
  - Send msgs to invite the user to attend a promotion activity.

<table>
<thead>
<tr>
<th></th>
<th>Online Test 1 2013.12.27 - 2014.1.3</th>
<th>Online Test 2 2014.1.24 - 2014.1.27</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Group name</strong></td>
<td>test group</td>
<td>control group</td>
</tr>
<tr>
<td>Group size</td>
<td>600K</td>
<td>200K</td>
</tr>
<tr>
<td>#Message read</td>
<td>345K</td>
<td>106K</td>
</tr>
<tr>
<td>Message read rate</td>
<td>57.50%</td>
<td>53.00%</td>
</tr>
<tr>
<td>#Message clicked</td>
<td>47584</td>
<td>7466</td>
</tr>
<tr>
<td>Message clicked rate</td>
<td>7.93%</td>
<td>3.73%</td>
</tr>
<tr>
<td>Lift_Ratio</td>
<td>196.87%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Summary

- Computational models for social tie analysis
  - Inferring social tie
  - Parasocial -> Reciprocal
  - Triadic closure
  - Cross-domain
- This is just a start for social tie analysis
  - How social tie influences user behaviors?
  - How social tie influences the network structure?
  - ...
Outline

- Introduction
- Friend Recommendation
- Content Recommendation
- Location Recommendation
- Summary
The Scope of Content Recommendation in the Tutorial

- Content may be manifested in diverse ways such as tweets, images, music, products, or videos
  - We do not assume that a item-feature matrix is available
  - User-item relations can be represented as a user-item matrix

- We only focus on collaborative filtering algorithms
  - Widely used
  - Promising performance in many real-world recommender systems

<table>
<thead>
<tr>
<th></th>
<th>i_1</th>
<th>i_2</th>
<th>i_3</th>
<th>i_4</th>
<th>i_5</th>
<th>i_6</th>
<th>i_7</th>
<th>i_8</th>
</tr>
</thead>
<tbody>
<tr>
<td>u_1</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u_2</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u_3</td>
<td>4</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>u_4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u_5</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>u_6</td>
<td>4</td>
<td>3</td>
<td></td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Challenges of Traditional Approaches

- **Data sparsity problem**
  - Content in social media is big but the available content for most individuals is very limited
  - The user-item matrix is extremely sparse with less than 1% observed entities

- **Cold-start users**
  - The number of entities per user follows a power-law distribution
  - Many users have no or very few entities

![Graph showing power-law distribution](image)
Opportunities from Social Media

- Social media provides additional sources for content recommendation
  - Social information and location information
  - Mitigating data sparsity problem

- We may make recommendations for cold-start users based on other information sources
  - Users’ preferences are similar to their networks
  - Reducing significantly the number of cold-start users
Content Recommendation

Content Recommendation with Social Networks

Location-aware Content Recommendation
Content Recommendation

Content Recommendation with Social Networks

Location-aware Content Recommendation
Why Social Networks

- Social networks can provide complementary information
  - Overlap between one’s similar users and her social network is less than 10% [Crandall et al., 2009]
- Users’ preferences are likely to similar to their networks
  - Homophily [McPherson et al., 2001]
  - Social influence [Marsden and Friedkin, 1993]
Most existing social recommender systems are CF-based methods

We can categorize social recommender systems based on their basic CF models
- Memory-based social recommender systems
- Model-based social recommender systems
Memory-based Social Recommendation

- It uses memory-based CF methods, especially user-oriented methods, as basic models

- It usually consists of two steps
  - Step 1: obtaining relevant users $N_i$ for user $i$,
  - Step 2: aggregating recommendations from $N_i$

- Different systems in this category provide different ways to obtain relevant users in Step 1
TidalTrust and MoleTrust

- TidalTrust only considers users at the shortest distance [Golbeck, 2005]
  - Efficient
  - High precision
  - Low recall

- MoleTrust considers raters up to a maximum-depth $d$ [Massa and Avesani, 2004]
  - Trade-off between precision and recall

\[
\hat{R}_{ij} = \bar{R}_i + \frac{\sum_{u_k \in N_i} S_{ik}(R_{kj} - \bar{R}_k)}{\sum_{u_k \in N_i} S_{ik}}
\]
In addition to distant users who have rated the target item, it also uses near friends who have rated similar items
  – Distant users on the exact target item
  – Close friends on similar items

It combines item-based recommendation and trust-based recommendation via random walk
Each random walk starts from a target user $u$ to seek rating score for item $i$

In step $k$ at user $v$:
- If $v$ has rated $i$, return $R_{vi}$
- With the probability $Q_{vik}$, stop random walk, select a similar item $j$ rated by $u$ and return $R_{vj}$
- With the probability $1 - Q_{vik}$, continue the random walk to a direct neighbor of $v$
Recommendation in Social Media
Continue?
Yes
Recommendation in Social Media

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Prediction = 4.67
Model-based Social Recommendation

- Model-based social recommender systems choose model-based CF methods as their basic models
  - Matrix factorization is widely chosen as the basic model

- There are three common ways to integrate social information under the matrix factorization framework [Tang et al., 2013]
  - Co-factorization methods
  - Ensemble methods
  - Regularization methods
Co-factorization Methods

- Co-factorization methods perform co-factorization on the user-item matrix $R$ and the user-user social matrix $S$

- SoRec [Ma et al., 2008]

\[ R_{ij} = U_i V_j^T \]
\[ S_{ij} = U_i Z_j^T \]
Ensemble Methods

- Ensemble methods combine recommendations for a user and her social network

- STE [Ma et al., 2009]

$$ R_{ij} = U_i V_j^T + \alpha \sum_{u_k \in N(u_i)} S_{ik} U_k V_j^T $$
Regularization Methods

- Regularization methods add a regularization term to force users’ preferences to be close to those of their social networks.

- SocialMF [Jamali and Ester, 2010]

\[
\min \sum_i \| U_i - \sum_{u_k \in N(u_i)} S_{ik} U_k \| 
\]

The average user preference of the social network of \( u_i \).
Heterogeneity

- Users trust their friends differently in different domains [Tang et al. 2012]
Circle-based Social Recommendation [Yang et al. 2012]

- Trust circle inference
  - $v$ is in inferred circle $c$ of $u$ iff $u$ connects to $v$ and both of them are interested in the category $c$

- SocialMF is applied to make recommendations for each circle
Distrust in Social Media [Tang et al., 2014]

- Distrust tends to be more noticeable and credible, and weighed more in decision making than trust.

- Distrust is not the negation of trust and has significant added value.
  - A small amount of distrust information can make remarkable improvement in link prediction.

Distrust and caution are the parents of security.
- Benjamin Franklin
Distrust in Social Recommendation [Victor et al., 2009]

- **Distrust as a filter**
  - Using distrust to filter out "unwanted" users in the recommendation processes

\[
\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{v \notin D} (r_{v,i} - \bar{r}_v) \times t_{u,v}}{\sum t_{u,v}}
\]

- **Distrust as a dissimilarity measure**

\[
\hat{r}_{u,i} = \bar{r}_u + \frac{\sum_{v \notin D} (r_{v,i} - \bar{r}_v) \times t_{u,v}}{\sum t_{u,v}} - \frac{\sum_{v \in D} (r_{v,i} - \bar{r}_v) \times d_{u,v}}{\sum d_{u,v}}
\]
Is Distrust Dissimilarity? [Tang et al., 2014]

- Distrust is not a dissimilarity measurement
  - CI: Commonly-rated Items
  - COSINE: Rating-cosine similarity
  - COSINE-CI: Rating-cosine similarity of commonly rated items

- Using distrust for recommendation is still an open challenge

<table>
<thead>
<tr>
<th></th>
<th>CI</th>
<th>COSINE</th>
<th>COSINE-CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distrust ($s_d$)</td>
<td>0.4994</td>
<td>0.0105</td>
<td>0.0142</td>
</tr>
<tr>
<td>Trust $s_t$</td>
<td>0.6792</td>
<td>0.0157</td>
<td>0.0166</td>
</tr>
<tr>
<td>Random Pairs ($s_r$)</td>
<td>0.1247</td>
<td>0.0027</td>
<td>0.0032</td>
</tr>
</tbody>
</table>
Content Recommendation

- Content Recommendation with Social Networks
- Location-aware Content Recommendation
Why Users’ Locations Matter?

- Users’ preferences may differ based on the user locations
  - New York Times has a very interesting visualization tool for Netflix rental patterns by zip code

http://www.nytimes.com/interactive/2010/01/10/nyregion/20100110-netflix-map.html?_r=0
Why Users’ Locations Matter?

- Users are more interested in content that is close to their current locations.
Why Users’ Locations Matter?

- When the recommended items are locations, users tend to travel a limited distance
  - 75% of users travel less than 50 mi
Location-aware Content Recommender Systems

- Location Distance Weighted Methods
  - Confounding effects

- User-partition Based Methods
  - Partition users based on their locations

- Item-partition Based Methods
  - Partition items based on their associated locations
Location Distance Weighted Methods [Yue et al. 2013]

- Geographically closed users are likely to share similar user preferences
  - Confounding

- Calculating location similarity as

\[ L_{uv} = \frac{1}{1 + \alpha \times \text{distance}(u, v)} \]

- Incorporating location similarity into user-oriented collaborative filtering

\[ R_{ui} = \sum_{v \in N_u} L_{uv} w_{uv} R_{vi} \]
User Partition Based Methods [Levandoski et al., 2012]

1. Partition ratings by user locations

Cell 1

Cell 2

Cell 3

2. Build collaborative filtering model for each cell using only ratings contained within the cell

Build Collaborative Filtering Model using:

<table>
<thead>
<tr>
<th>User</th>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

Build Collaborative Filtering Model using:

<table>
<thead>
<tr>
<th>User</th>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

Build Collaborative Filtering Model using:

<table>
<thead>
<tr>
<th>User</th>
<th>Item</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>5</td>
</tr>
</tbody>
</table>

3. Generate recommendations using collaborative filtering using the model of the cell containing the target user

Target user

Recommendation List
User Partition Structure

- Adaptive Pyramid Structure
  - 7th Ave, New York, NY, USA

- Two main goals:
  - Locality
  - Scalability

Smaller cells $\rightarrow$ more “localized” answers
Item-partition based Methods

• Partition items based on their associated locations

• Penalizing the item based on its distance from the user

• Recommending items within a certain distance
References


References


References


Outline

Introduction
Friend Recommendation
Content Recommendation
Location Recommendation
Summary
Location-based Social Networks [Gao and Liu, 2014]

- Location-Based Social Networking Sites
  - Foursquare, Facebook Places, Yelp
Information in LBSNs

Content Layer
User-Generated Content

Social Layer
Social Friendships

Geographical Layer
Check-in POIs

Timeline
Time Stamps
Challenges for Location Recommendation

- Geographical Properties of Social Connections
  - Geographical Distance
  - Social Connections

- Temporal Cyclic Patterns of Geographical Check-ins
  - Going to restaurant around noon
  - Watching movie in a theater during the weekend

- Content information could be very important
Location recommender systems can be divided into three groups according to the information used:

- **Geo-temporal Location Recommendation**
- **Geo-content Location Recommendation**
- **Geo-social Location Recommendation**
Location Recommendation in Social Media

- Geo-social Location Recommendation
- Geo-temporal Location Recommendation
- Geo-content Location Recommendation
Geographical Properties of Social Connections

- There is a strong correlation between friendship and trajectory similarity in LBSNs
  
  ![Graph showing correlation between friendship and trajectory similarity](image)

  [Cho et al., 2011]

- Nearby friends have a much higher probability to share common locations
  
  ![Graph showing mean common location ratio](image)

  [Mao et al., 2010]
Friend-based Methods [Mao et al., 2010]

- Friend-based Collaborative Filtering: FCF

\[ \hat{r}_{i,j} = \frac{\sum_{u_k \in U_i} r_{k,j} w_{i,k}}{\sum_{u_k \in U_i} w_{i,k}} \]

- Geo-Measured FCF: GM-FCF
  - Assuming a power-law relation between trajectory similarity \( y \) and geographical distance \( x \)

\[ y = \alpha x^\beta \]

- Similarity is computed as

\[ w_{i,k} = \frac{y(x = d(u_i, u_k), \alpha, \beta)}{\sum_{u_k \in U_i} y(x = d(u_i, u_k), \alpha, \beta)} \]
Preference and Friend based methods

- A fusion model: USG [Mao et al., 2011]
  - The probability score of i-th user at j-th location is
    \[
    S_{i,j} = (1 - \alpha - \beta)S_{i,j}^u + \alpha S_{i,j}^s + \beta S_{i,j}^g
    \]

- A Social-Historical Model: SHM [Gao et al., 2012a]
  - Users’ historical information is modeled by Hierarchical PitmanYor process
    \[
    P_{i,j}^{SH} = \alpha P_{i,j} + (1 - \alpha) \sum_{u_k \in N_i} w_{i,k} P_{k,j}
    \]
Some Observations for Geo-social Location Recommendation

- Social information can consistently improve the recommendation performance, however, the improvement is very limited.

[Mao et al., 2011] [Gao et al., 2012a]
Geo-social Circles [Gao et al., 2012b]

- Friends with long distance share a small number of commonly visited locations

- Non-friends with short distance share a large number of commonly visited locations

- Users are segmented into four geo-social circles

<table>
<thead>
<tr>
<th>F</th>
<th>Geo-Social Circles</th>
<th>$\overline{F}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{FD}$:</td>
<td>$\overline{S}_{FD}$: Local Friends</td>
<td>$S_{FD}$: Local Non-friends</td>
</tr>
<tr>
<td>$D$</td>
<td>$S_{FD}$: Distant Friends</td>
<td>$S_{FD}$: Distant Non-friends</td>
</tr>
</tbody>
</table>
A framework is proposed to address cold-start problem in location recommendation based on geo-social circles.

\[
P_u^t(l) = \Phi_1 P_u^t(l|S_{FD}) + \Phi_2 P_u^t(l|S_{FD}) + \Phi_3 P_u^t(l|S_{FD}) + \Phi_4 P_u^t(l|S_{FD}).
\]

“A New Check-in” at location l

User U

L1

L2

L3

L4
Observations about Geo-social Circles

- Local friends are more important than distant friends
- Distance friends contain more additional information than local friends when combining with local non-friends
- These four geo-social circles contain complementary information although their contributions differ

<table>
<thead>
<tr>
<th>Methods</th>
<th>Top-1</th>
<th>Top-2</th>
<th>Top-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{FD}$</td>
<td>6.51%</td>
<td>8.31%</td>
<td>9.32%</td>
</tr>
<tr>
<td>$S_{FD}$</td>
<td>3.65%</td>
<td>4.75%</td>
<td>5.34%</td>
</tr>
<tr>
<td>$S_{FD}$</td>
<td>18.37%</td>
<td>24.10%</td>
<td>27.34%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD}$</td>
<td>18.62%</td>
<td>24.44%</td>
<td>27.79%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD}$</td>
<td>19.01%</td>
<td>24.95%</td>
<td>28.35%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD}$</td>
<td>8.33%</td>
<td>10.79%</td>
<td>12.23%</td>
</tr>
<tr>
<td>$S_{FD} \cup S_{FD} \cup S_{FD}$</td>
<td>19.21%</td>
<td>25.19%</td>
<td>28.69%</td>
</tr>
</tbody>
</table>
Location Recommendation in Social Media

- Geo-social Location Recommendation
- Geo-temporal Location Recommendation
- Geo-content Location Recommendation
Why Temporal Information Matters?

- Human movement exhibits strong temporal cyclic patterns
  - Days of the week patterns [Cho et al., 2011]
  - Hours of the day patterns [Gao et al., 2013a]
Location Recommendation with Time Preference: UT
[Yuan et al., 2013]

- Splitting data into 24 slots based on hours
  - Nov. 6 2012, 10:30 → 10

- Introducing time dimension into user-location matrix $c$
  - $c_{u,l} \rightarrow c_{u,t,l}$

- Leveraging time factor when
  - Computing the similarities between users over time

\[
W_{u,v}^{(t)} = \frac{\sum_t \sum_l c_{u,t,l} c_{v,t,l}}{\sqrt{\sum_t \sum_l c_{u,t,l}^2} \sqrt{\sum_t \sum_l c_{v,t,l}^2}}
\]

- Making predictions

\[
\hat{c}_{u,t,l} = \frac{\sum_v W_{u,v}^{(t)} c_{v,t,l}}{\sum_v W_{u,v}^{(t)}}
\]
Enhancing UT by Smoothing

- Data in each slot becomes even sparser after splitting
- Check-in behaviors of users at different time are correlated
- Smoothing $c_{u,t,l}$ based on the similarity between different time slots

\[
\tilde{c}_{u,t,l} = \sum_{t' = 1}^{T} \frac{\rho_{t,t'}}{\sum_{t'' = 1}^{T} \rho_{t,t''}} c_{u,t',l}
\]
Enhancing UT by Location Popularity

- The popularity of a location varies over time
  - A restaurant is more popular around noon and evening
- Location popularity is calculated as

\[
P_t(l) = \beta \frac{|CI_{l,t}|}{\sum_{l' \in L} |CI_{l',t}|} + (1 - \beta) \frac{|CI_l|}{\sum_{l' \in L} |CI_{l'}|}
\]

Number of Check-ins at \(l\) at time \(t\)

Number of Check-ins at \(l\)
Location Recommendation with Temporal Effects [Gao et al., 2013]

- One user’s daily check-in activity w.r.t. his top 5 frequently visited locations

![Graph showing temporal effects]

- Temporal Non-uniformness
  - A user presents different check-in preferences at different hours of the day

- Temporal Consecutiveness
  - A user presents similar check-in preferences at nearby hours of the day
Modeling Temporal Non-uniformness

- A user presents different check-in preferences at different hours of a day

\[
\begin{align*}
\min_{U_i \geq 0, L_j \geq 0} & \sum_{i} \sum_{j} Y_{i,j} (C_{i,j} - U_i L_j^T)^2 \\
\min_{U_i \geq 0, L_j \geq 0} & \sum_{t=1}^{24} \sum_{i} \sum_{j} Y_{i,j}^t (C_{i,j}^t - U_i^t L_j^T)^2
\end{align*}
\]
A user presents similar check-in preferences at nearby hour of the day

\[ \min_{U \geq 0} \sum_{t=1}^{T} \sum_{i=1}^{m} \psi_i(t, t-1) \| U_t(i,:) - U_{t-1}(i,:) \|_F^2 \]

\[ \psi_i(t, t-1) = \frac{C_t(i,:) \cdot C_{t-1}(i,:)}{\sqrt{\sum_j C_t^2(i,:) \sqrt{\sum_j C_{t-1}^2(i,:)}}} \]
Framework of Location Recommendation with Temporal Effects

Unobserved Check-ins

T=24

Approximated Check-in Preference
Location Recommendation in Social Media

- Geo-social Location Recommendation
- Geo-temporal Location Recommendation
- Geo-content Location Recommendation
Content in LBSNs

- Content in LBSNs is pervasively available
  - Tags, tips or comments

- Content contains semantic words that reflect a user’s interested topics and the location property
  - “Chinese” and “Spicy”

- Content can reflect users’ preferences
  - “all great”
Why Sentiment in Content is Important?

- Ratings in traditional recommendation can capture user preferences
  - Like/dislike, voting scores from 1 to 5

- Check-in behavior represents users’ habitual behavior and may not be sufficient to reflect users’ preferences
  - High check-in frequencies may represent positive opinions
  - Fewer checked locations are not necessarily less favored

- Sentiment extracted from content contains more precise information about a user’s preference on a location
  - In addition to positive feedback, there could also be negative feedback from content
Sentiment-enhanced Location Recommendation [Yang et al., 2013]

- Extracting check-in preferences from check-in data
- Extracting sentiment preferences from content
- Combining check-in preferences and sentiment preferences
- Performing traditional CF based on the combined preferences
Preference Extraction from Check-ins

- Check-in frequencies can reflect users’ preferences
  - Users prefer those locations with high check-in frequencies

- Mapping frequencies to five-point preferences
  - Check-in frequencies follow the power law distribution

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Preference Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2 (Fair)</td>
</tr>
<tr>
<td>2</td>
<td>3 (Good)</td>
</tr>
<tr>
<td>3</td>
<td>4 (Very Good)</td>
</tr>
<tr>
<td>&gt;=4</td>
<td>5 (Excellent)</td>
</tr>
</tbody>
</table>

- Constructing a check-in preference matrix $P_c$
Preference Extraction from Content

- Sentiment extracted from content reflects user’s preference on a location
- Mapping sentiment scores to five-point preferences
  - Sentiment scores are highly centralized around 0
  - A slight bias towards positive sentiment
- Constructing a sentiment preference matrix $P_s$

<table>
<thead>
<tr>
<th>Sentiment Scores</th>
<th>Preference Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>[-1,-0.05]</td>
<td>1</td>
</tr>
<tr>
<td>(0.05,-0.01]</td>
<td>2</td>
</tr>
<tr>
<td>(-0.01,0.01]</td>
<td>3</td>
</tr>
<tr>
<td>[0.01,0.05]</td>
<td>4</td>
</tr>
<tr>
<td>[0.01,1]</td>
<td>5</td>
</tr>
</tbody>
</table>
Combining the check-in preference matrix and the sentiment preference matrix

- Sentiment preference has a bigger impact for one-time check-in locations
- Sentiment preference has some impact for multi-time check-in locations

\[ P_{final} = P_C + \text{sgn}(P_C - P_S) \cdot H(|P_C - P_S| - 2) \]
Geographical topics are discovered from LBSNs [Yin et al., 2011]

- Assigning semantic topics to locations
- Reflecting users’ interests
- Connecting users and locations in the semantic level
Topic-aware Location Recommendation [Liu and Xiong, 2013]

- Building an aggregated LDA model to discover geographical topics
  - User interest topic distribution $\theta_i$
  - Location topic distribution $\pi_j$

- Defining topic and location influence index

$$TL_{ij} = \alpha(1 - D_{JS}(\theta_i, \pi_j)) + (1 - \alpha)P_j$$

- Jensen-Shannon Divergence
- Location Popularity

- Modeling users check-in behaviors as

$$c_{ij} = TL_{ij}U_i^T C_j$$
References


References


Outline

Introduction
Friend Recommendation
Content Recommendation
Location Recommendation
Summary
Recommendation in Social Media

- Friend Recommendation
- Content Recommendation
- Location Recommendation
Recommendation in Social Media

- Friend Recommendation
- Content Recommendation
- Location Recommendation

- Inferring Social Ties
- Reciprocity
- Triadic closure
- Cross-community
Recommendation in Social Media

- Friend Recommendation
  - Inferring Social Ties
    - Reciprocity
    - Triadic closure
    - Cross-community
  - Challenges and Opportunities
  - Content Recommendation with Social Networks
  - Location-aware Content Recommendation

- Content Recommendation

- Location Recommendation

- Challenges and Opportunities
  - Content Recommendation with Social Networks
  - Location-aware Content Recommendation
Recommendation in Social Media

- Friend Recommendation
- Content Recommendation
- Location Recommendation

- Challenges and Opportunities
- Content Recommendation with Social Networks
- Location-aware Content Recommendation

- Inferring Social Ties
  - Reciprocity
  - Triadic closure
  - Cross-community

- Location-based Social Networks
  - Geo-social Location Recommendation
  - Geo-temporal Location Recommendation
  - Geo-content Location Recommendation
Recommending with Cross-Media Data

- Users usually join multiple social media sites
  [Zafarani and Liu, 2014]
  - More than 97% of users have joined at most 5 sites
  - Users exist on as many as 16 sites

- A new user on one site might have existed on other sites for a long time
  - Cross media data can mitigate data sparsity problem
  - Cross media data can reduce cold-start users
Deep Learning in Recommendation

- Deep learning has been proven to be effective in various domains
  - Pattern recognition and natural language processing

- Recently deep convolutional neural networks is used to predict latent factors from music audio for music recommendation [VanDeOord et al., 2013]
  - A content-based method without data sparsity problem in collaborative filtering
  - Viable for recommending new and unpopular music

- How to apply deep learning with rich social media data is still an open issue
Privacy-preserving Recommendation

- Recommender systems in social media may utilize sensitive information from users to produce better recommendations
  - Users’ locations in location-aware content recommendation
  - Social networks in social recommendation
  - Check-in data in location recommendation

- New privacy threats are introduced by recommender systems in social media [Jeckmans et al., 2013]
  - The privacy of social relations
  - The privacy of their locations
References


Collaborators: John Hopcroft, Jon Kleinberg, Chenhao Tan (Cornell)
Jiawei Han and Chi Wang (UIUC)
Tiancheng Lou (Google) Jimeng Sun (IBM)
Wei Chen, Ming Zhou, Long Jiang (Microsoft)
Jing Zhang, Zhanpeng Fang, Zi Yang, Sen Wu, Jia Jia (THU)

Jie Tang, KEG, Tsinghua U,
Download all data & Codes,
http://keg.cs.tsinghua.edu.cn/jietang
http://arnetminer.org/download
Acknowledgements

- Projects are partially supported by National Science Foundation, Army Research Office and The Office of Naval Research

- Members of Data Mining and Machine Learning Lab at ASU provided valuable feedback and suggestions