Recommender Systems in Social Media
Traditional Recommender Systems

- Collaborative Filtering

User1 and User2 both like item 2
People like item 2 also like item 3
People like item 2 also like item 1
Traditional Recommender Systems

• Problem Definition

- Have seen before or not
- Like or dislike?
- Numerical rating

<table>
<thead>
<tr>
<th>User</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
</tr>
</thead>
<tbody>
<tr>
<td>User1</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>User2</td>
<td>?</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
</tr>
</thead>
<tbody>
<tr>
<td>User1</td>
<td>-1</td>
<td>+1</td>
<td>?</td>
</tr>
<tr>
<td>User2</td>
<td>?</td>
<td>+1</td>
<td>-1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>User</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
</tr>
</thead>
<tbody>
<tr>
<td>User1</td>
<td>90</td>
<td>80</td>
<td>?</td>
</tr>
<tr>
<td>User2</td>
<td>?</td>
<td>80</td>
<td>70</td>
</tr>
</tbody>
</table>
Recommender System in Social Media

• Different types of nodes and edges

Social Network
Follower-followee
Friends

Tag Semantic
Relatedness / Correlation

Item Similarities based on profiles like textual description/categories
Recommender System in Social Media

- Node and edges contain rich metadata.
Outline

• Personalized Tag Recommendation [WWW’12, KDD’12, IJCAI’13]
  • Random Walks on Heterogeneous Network

• Personalized Tweet Re-ranking [WSDM’13]
  • Matrix Factorization on Heterogeneous Network

• Learning to Annotate Tweets [WWW’13, ICDE’14]
  • Matrix Factorization on Heterogeneous Network
Incorporating Heterogeneous Information for Personalized Tag Recommendation in Social Tagging Systems
Incorporating Heterogeneous Information for Personalized Tag Recommendation in Social Tagging Systems
What is a social tagging system

Search for KDD 2012
What is a social tagging system

How to format a JSON date?

I'm taking my first crack at Ajax with jQuery. I'm getting my data onto my page, but I'm having some trouble with the JSON data that is returned for Date data types. Basically, I'm getting a string back that looks like this:

```
/Date(1224034300000)/
```

From someone totally new to JSON - How do I format this to a short date format? Should this be handled somewhere in the jQuery code? I've tried the `$.datepicker.formatDate()` without any success.

FYI: Here's the solution I came up with using a combination of the answers here.

```javascript
function getMismatch(id) {
    $.getJSON("Main.aspx?Callback=GetMismatch",
    {
    MismatchId: id
    },
    function(result) {
        $("#AuthMerchId").text(result.OrganizationMerchantId);
        $("#StllMerchId").text(result.SettlementMerchantId);
        $("#CreateDate").text(formatJSONDate(Date(result.AppendDts)));
        $("#ExpireDate").text(formatJSONDate(Date(result.ExpiresDts)));
        $("#PersistDate").text(formatJSONDate(Date(result.LastUpdateDts)));
    });
}
```
What is a social tagging system
Incorporating **Heterogeneous Information** for Personalized Tag Recommendation in Social Tagging Systems
What is “Heterogeneous Information”

- **Social Network**: Follower-followee, Friends
- **Tag Semantic Relatedness / Correlation**
- **Item Similarities** based on profiles like textual description/categories
Incorporating Heterogeneous Information for Personalized Tag Recommendation in Social Tagging Systems
What is personalized tag recommendation

Personalized Tag Recommendation
Why personalized tag recommendation

Users have their own way to organize items

Filter items by tag

TAGS

- blog: 2
- technology: 2
- development: 1
- web2.0: 1
- code: 1
- business: 1
- twitter: 1
- engineering: 1
- tech: 1
- news: 1
- review: 1
Our Contribution

Incorporating Heterogeneous Information for Personalized Tag Recommendation in Social Tagging Systems
Why “**heterogeneous Information**” matters

• Users in a social network may influence each other by sharing some annotated items.

• Semantically related tags may co-occur to describe an item.

• Items that have similar contents may be annotated with the same tag.

Alleviate data sparsity!
Basic solution for **tag recommendation**

- Random walk with restart

```
Start here
```

```
and start here
```

```
Users
```

```
Tags
```

```
Items
```
What about item recommendation

- Random walk with restart
  Start here
  and start here

- Users
- Tags
- Items
What about user recommendation

• Random walk with restart
Algorithm

- OptRank (optimized random walk with restart)

\[
\begin{pmatrix}
    p_U \\
    p_T \\
    p_I
\end{pmatrix}^{(t+1)} = (1 - \alpha) A \begin{pmatrix}
    p_U \\
    p_T \\
    p_I
\end{pmatrix}^{(t)} + \alpha \begin{pmatrix}
    \bar{q}_U \\
    0 \\
    \bar{q}_I
\end{pmatrix}
\]

\[A = \begin{bmatrix}
    A_{UU} & A_{UT} & A_{UI} \\
    A_{TU} & A_{TT} & A_{TI} \\
    A_{IU} & A_{IT} & A_{II}
\end{bmatrix}\]
The first challenge

- Where to restart? 50% at $u_1$ and 50% at $i_1$?
The second challenge

- Different types of edges are measured in different scale
- The importance of differs
  - which type of edge is more important?

$$\begin{align*}
A &= \begin{bmatrix}
A_{UU} & A_{UT} & A_{UI} \\
A_{TU} & A_{TT} & A_{TI} \\
A_{IU} & A_{IT} & A_{II}
\end{bmatrix} \\
\overline{A} &= \begin{bmatrix}
A_{UU}D_{U}^{-1} & A_{UT}D_{T}^{-1} & A_{UI}D_{I}^{-1} \\
A_{TU}D_{U}^{-1} & A_{TT}D_{T}^{-1} & A_{TI}D_{I}^{-1} \\
A_{IU}D_{U}^{-1} & A_{IT}D_{T}^{-1} & A_{II}D_{I}^{-1}
\end{bmatrix}
\end{align*}$$

Not comparable
Nodes and edges with features

• Each node is represented by a feature vector

\[ \mathbf{q}_I(i) = f_{node}(\xi_I^T \mathbf{Y}_I) \]

\[ \mathbf{q}_U(u) = f_{node}(\xi_U^T \mathbf{Y}_U) \]

• Each edge is represented by a feature vector

\[ \mathbf{A}(u, v) = f_{edge}(\theta^T \mathbf{X}(u, v)) \]
Optimization

• Loss Function

\[
\max_{\theta, \xi} AUC(\theta, \xi) = \frac{\sum_{i \in PT} \sum_{j \in NT} \mathbb{I}(p_T(i) - p_T(j))}{|PT||NT|}
\]

Use a differentiable version

\[
\min_{\theta, \xi} J(\theta, \xi) = \frac{1}{m} \sum_{k=1}^{m} \sum_{i \in PT_k} \sum_{j \in NT_k} S(p_T(j) - p_T(i)) \frac{1}{|PT_k||NT_k|}
\]

• Stochastic gradient descent

\[
\theta^{(t+1)} = \theta^{(t)} - lr \frac{\partial J_k(\theta^{(t)}, \xi^{(t)})}{\partial \theta}
\]

\[
\xi^{(t+1)} = \xi^{(t)} - lr \frac{\partial J_k(\theta^{(t)}, \xi^{(t)})}{\partial \xi}
\]
Datasets

• Delicious
  • 437k posts, 69k items, 40k tags and 2k users
  • 151k item-item relation, 197k tag-tag relation, 15k user-user relation

• Last.fm
  • 24k posts, 12k users, 9k tags and 2k users
  • 25k user-user relation
Results

• Naïve Random Walk with Restart
• FolkRank [Jaschke, PKDD2007]
Summary

• Data as a graph
• Define node/edge features depending on the application
• Optimization
• Random walk
Efficient Personalized PageRank

• Yasuhiro Fujiwara, Makoto Nakatsuji, Takeshi Yamamuro, Hiroaki Shiokawa, Makoto Onizuka: "Efficient Personalized PageRank with Accuracy Assurance", KDD2012

• Yasuhiro Fujiwara, Makoto Nakatsuji, Makoto Onizuka, Masaru Kitsuregawa: "Fast and Exact Top-k Search for Random Walk with Restart", VLDB 2012
Outline

• Personalized Tag Recommendation [WWW’12, KDD’12, IJCAI’13]
  • Random Walks on Heterogeneous Network

• Personalized Tweet Re-ranking [WSDM’13]
  • Matrix Factorization on Heterogeneous Network

• Learning to Annotate Tweets [WWW’13, ICDE’14]
  • Matrix Factorization on Heterogeneous Network
What is Tweet Re-ranking?

- Boring tweets come first
- Latest tweets come first
- Interesting tweets come first

Examples:
- Boring: Mikio L. Braun @mikiobraun: Funny that I almost exclusively use the winking ;) instead of :) although I never wink in person.
- Boring: IBMResearch @IBMResearch: Shhh... This Inventive Mind is in deep thought. bmlr.co /Zb2FVtbrbK2I #innovation #think20
- Interesting: NetEvolution @NetEvolution: Facebook announces 'Graph Search' to search its connections - MW buff.ly/XbppgU
- Interesting: Vincent Granville @vitalvinge: Using TeXmacs as an interface for R: (This article was first published on MLT thinks, and kindly contributed ... bit.ly/ZT2LDQ
- Personalized Tweet Re-ranking: Mikio L. Braun @mikiobraun: Funny that I almost exclusively use the winking ;) instead of :) although I never wink in person.
- Personalized Tweet Re-ranking: IBMResearch @IBMResearch: Shhh... This Inventive Mind is in deep thought. bmlr.co /Zb2FVtbrbK2I #innovation #think20
Why Re-ranking Matters?

• If you follow many users
  • handle overwhelmed tweets
• If you did not log in for a long time
  • thousands of tweets are missed
• If you are using a mobile phone with limited bandwidth
  • loading pages of tweets takes time
• Solution
  • Personalized Tweet Re-ranking, i.e., most interesting tweets come first
How to measure interestingness?

• Retweet Behavior
  – Retweet => Interesting
  – Adopted in this paper

• Other options
  – Comments => Interesting
  – Favorite => Interesting
  – Clicks / Stay Time => Interesting
How to model the data?
How to model the data?

• **Node Weight**
  – whether a user is willing to retweet
  – whether a publisher is likely to be retweeted
  – whether tweet i has a high quality

• **Edge Weight**
  – whether user u and publisher p are close friends
  – whether tweet i is interesting to user u
Framework

• Ranking Function

\[ \hat{r}_{upi} = \sum_{m \in \{u,p,i\}} f_m + \sum_{m \in \{p,i\}} f_{um} + \sum_{m \in \{p,i\}} g_u^T g_m \]

Node Weights  Edge Weights  Factorization on ID features

• Loss Function
  – Logistic loss function
  – Area Under the Curve (AUC)
Experiments

• Dataset
  – Crawling Twitter according to breadth-first strategy from April to June, 2012

<table>
<thead>
<tr>
<th>Table 1: Dataset Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
</tr>
<tr>
<td>28,420</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Overlap between Test Set and Train Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
</tr>
<tr>
<td>93.6%</td>
</tr>
</tbody>
</table>

• Baselines
Experimental Results

- Personalized methods are better than non-personalized methods
- Factorization on tweet ID will fail
- Optimizing AUC is slightly better than optimizing logistic regression
Experimental Results

User trust is the most effective feature
Outline

• Personalized Tag Recommendation [WWW’12, KDD’12, IJCAI’13]
  – Random Walks on Heterogeneous Network

• Personalized Tweet Re-ranking [WSDM’13]
  – Matrix Factorization on Heterogeneous Network

• Learning to Annotate Tweets [WWW’13, ICDE’14]
  – Matrix Factorization on Heterogeneous Network
What is a hashtag?

• #Election2012

Figure 1: Tweets containing #Election2012
Annotate Tweet Automatically

- Organized Tweets into Hashtag-based Topics
- Event Detection and Tracking
- Improve Search and Advertising
Is it possible to infer hashtags from tweets?

- Indicator from links and mentions

Table 1: Top-3 Hashtags for popular links and mentions

<table>
<thead>
<tr>
<th></th>
<th>Amazon</th>
<th>BBC Tech.</th>
<th>@BarackObama</th>
<th>@JLin7</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Win</td>
<td></td>
<td>#Apple</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#VideoGame</td>
<td>#Google</td>
<td></td>
<td>#Obama2012</td>
<td>#Knicks</td>
</tr>
<tr>
<td>#GiftCard</td>
<td>#Twitter</td>
<td></td>
<td>#Iran</td>
<td>#Linsanity</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>#Israel</td>
<td>#JeremyLin</td>
</tr>
</tbody>
</table>
Is it possible to infer hashtags from tweets?

• Inferring hashtag by locations

Fig. 4. Comparison of Hashtag Distributions of New York and Toronto
Is it possible to infer hashtags from tweets?

- Inferring hashtags from user accounts

<table>
<thead>
<tr>
<th>SAP</th>
<th>IBMBigData</th>
</tr>
</thead>
<tbody>
<tr>
<td>#SAP</td>
<td>#bigdata</td>
</tr>
<tr>
<td>#HANA</td>
<td>#hadoop</td>
</tr>
<tr>
<td>#BI</td>
<td>#analytics</td>
</tr>
</tbody>
</table>

(a) Users

- Inferring hashtags from time

Fig. 4. Comparison of Hashtag Distributions of New York and Toronto
Framework

- **Node features**
  - For example: user influence, hashtag popularity, tweet quality
- **Edge features**
  - For example: social relation, content relevance, user relevance
Experiments

• Dataset

Table 1: Data Statistics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#User</th>
<th>#Social Relation</th>
<th>#Tweet</th>
<th>#Hashtag</th>
<th>#Links</th>
<th>#Mention</th>
<th>#Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Month-Week</td>
<td>91,896</td>
<td>1,092,634</td>
<td>1,889,186</td>
<td>43,678</td>
<td>76,559</td>
<td>105,246</td>
<td>15,454</td>
</tr>
<tr>
<td>Week-Day</td>
<td>56,968</td>
<td>584,018</td>
<td>465,373</td>
<td>20,137</td>
<td>23,931</td>
<td>15,108</td>
<td>10,647</td>
</tr>
</tbody>
</table>

• Performance

Fig. 7. Comparison of Different Models
Summary

• Application
  – Personalized Tag Recommendation
  – Personalized Tweet Re-ranking
  – Learning to Annotate Tweets

• Techniques
  – Random walks on heterogeneous network
  – Matrix factorization on heterogeneous network
References

• Wei Feng, Jianyong Wang, Wei Zhang. We Can Learn Your #Hashtags: Connecting Tweets to Explicit Topics. ICDE 2014
• Wei Zhang, Wei Feng, Jianyong Wang. Integrating Semantic Relatedness and Words’ Intrinsic Features for Keyword Extraction. IJCAI 2013
• Wei Zhang, Jianyong Wang, Wei Feng. Combining Latent factor Model with Location Features for Event-based Group Recommendation. KDD 2013
• Wei Feng, Jianyong Wang. Learning to Annotate Tweets with Crowd Wisdom, WWW 2013
• Wei Feng, Jianyong Wang. Retweet or not? Personalized Tweet Re-ranking, WSDM 2013
• Wei Feng, Jianyong Wang. Incorporating Heterogeneous Information for Personalized Tag Recommendation in Social Tagging Systems, KDD 2012
Thank you!