Learning with Social Media

Tom Chao Zhou @Thesis Defense

Thesis Committee:
Prof. Yu Xu Jeffrey (Chair)
Prof. Zhang Sheng Yu (Committee Member)
Prof. Yang Qiang (External Examiner)

Supervisors:
Prof. Irwin King
Prof. Michael R. Lyu
Learning with Social Media

Introduction

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Learning with Social Media

Introduction

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Social Media

• What is Social Media?
  – Create, share, exchange; virtual communities

• Some Data
  – 45 million reviews in a travel forum TripAdvisor [Source]
  – 218 million questions solved in Baidu Knows [Source]
  – Twitter processed one billion tweets in Dec 2009, averages almost 40 million tweets per day [Source]
  – Time spent on social media in US: 88 billion minutes in July 2011, 121 billion minutes in July 2012 [Source]
Examples of Social Media

• Rating System

*amazon.com*  
America’s largest online retailer

*Taobao.com*  
The largest C2C website in China, over 2 billion products

*IMDb*  
The biggest movie site on the planet, over 1,424,139 movies and TV episodes
Examples of Social Media

• Social Tagging System

  del.icio.us

  The largest social bookmarking website

  social bookmarking

  Flickr

  The best online photo management and sharing application in the world
Examples of Social Media

- Online Forum
Examples of Social Media

• Community-based Question Answering

  Yahoo! Answers

  10 questions and answers are posted per second

  Baidu 知道

  218 million questions have been solved

  Quora

  A popular website with many experts and high quality answers
Challenges in Social Media

- Astronomical growth of data in Social Media
- Huge, diverse and dynamic
- Drowning in information, information overload
Objective of Thesis

• Establish automatic and scalable models to help social media users find their information needs more effectively
Objective of Thesis

- Modeling **users’ interests** with respect to their behavior, and recommending **items** or **users** they may be interested in
  - Chapter 3, 4
- Understanding **items’ characteristics**, and grouping **items** that are semantically related for better addressing users’ information needs
  - Chapter 5, 6
Structure of Thesis

- Participant
- Social Media
- Consumption Goods

- User
  - Item
    - Chap. 3
  - User
    - Chap. 4
  - Item
    - Chap. 5
  - Characteristic
    - Chap. 6
Learning with Social Media

Introduction

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Recommender Systems

- **Memory-based algorithms**
  - User-based
  - Item-based

- **Similarity methods**
  - Pearson correlation coefficient (PCC)
  - Vector space similarity (VSS)

- **Disadvantage of memory-based approaches**
  - Recommendation performances deteriorate when the rating data is sparse
Recommendation Systems

- Model-based algorithms
  - Clustering methods
  - Matrix factorization methods
- Disadvantage of traditional model-based approaches
  - Only use the user-item rating matrix, ignore other user behavior
  - Suffer the problem of data sparsity
Machine Learning

• Whether the training data is available
  • Yes? Supervised learning
    – Naive Bayes, support vector machines
  • Some? Semi-supervised learning
    – Co-training, graph-based approach
  • No? Unsupervised learning
    – Clustering, Latent Dirichlet Allocation
Information Retrieval

- **Information Retrieval Models**
  - Seek an optimal ranking function
- **Vector Space Model**
  - Weighting (TF-IDF)
- **Probabilistic Model and Language Model**
  - Binary independence model, query likelihood model
- **Translation Model**
  - Originated from machine translation
  - Solve the lexical gap problem
Techniques Employed

- Recommender Systems
- Information Retrieval
- Machine Learning

Chapter 3
Chapter 4
Chapter 5
Chapter 6
Learning with Social Media

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Structure of Thesis

- Participant
- Social Media
- Consumption Goods

User
- Item
  - Chap. 3

User
- Item
  - Chap. 4

Item
- Characteristic
  - Chap. 5
  - Chap. 6
# A Toy Example

<table>
<thead>
<tr>
<th></th>
<th>The Godfather</th>
<th>Inception</th>
<th>Forrest Gump</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex</td>
<td>4</td>
<td>?</td>
<td>5</td>
</tr>
<tr>
<td>Bob</td>
<td>4</td>
<td>2</td>
<td>?</td>
</tr>
<tr>
<td>Tom</td>
<td>?</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

1: Strong dislike, 2: Dislike, 3: It’s OK, 4: Like, 5: Strong like
Challenge

- Rating matrix is very sparse, density of ratings in commercial recommender system is less than 1%
- Performance deteriorates when rating matrix becomes sparse
Problem

Task: Predicting the missing values

User-item rating matrix

<table>
<thead>
<tr>
<th></th>
<th>i₁</th>
<th>i₂</th>
<th>i₃</th>
<th>i₄</th>
<th>i₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>u₁</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>u₂</td>
<td>?</td>
<td>4</td>
<td>?</td>
<td>4</td>
<td>?</td>
</tr>
<tr>
<td>u₃</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>u₄</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>u₅</td>
<td>?</td>
<td>5</td>
<td>?</td>
<td>4</td>
<td>?</td>
</tr>
</tbody>
</table>

Fact:
Ratings reflect users’ preferences

Challenge:
Rating matrix is very sparse, only use rating information not enough

Thought:
Whether there exists contextual information that can also reflect users’ judgments?

How can we utilize that kind of contextual information to improve the prediction quality?
Motivation

- Social tagging is to collaboratively creating and managing tags to **annotate** and **categorize** content.
- Tags can represent users’ **judgments** and **interests** about Web contents quite accurately.
To improve the recommendation quality and tackle the data sparsity problem, fuse tagging and rating information together.
Intuition of Matrix Factorization

- $M = U^T \star V, M \in \mathbb{R}^{m \times n}, U \in \mathbb{R}^{l \times m}, V \in \mathbb{R}^{l \times n}, l \ll (m, n)$

- Physical meaning of each row in $U$ and $V$ is a latent semantic dimension
- E.g., action, comedy, if $M$ is a user-movie rating matrix
User-Item Rating Matrix Factorization

Conditional distributions over the observed

\[ p(R|U, V, \sigma^2_R) = \prod_{i=1}^{m} \prod_{j=1}^{n} [\mathcal{N}(r_{ij}|g(U_i^T V_j), \sigma^2_R)]^{R_{ij}} \]

- **U**: user latent feature matrix.
- **V**: item latent feature matrix.
- **U_i^T V_j**: predicted rating (user i to item j).

Zero-mean spherical Gaussian priors are placed on the user latent feature matrix and the item latent feature matrix

\[
\begin{align*}
p(U|\sigma^2_U) & = \prod_{i=1}^{m} \mathcal{N}(U_i|0, \sigma^2_U I) \\
p(V|\sigma^2_V) & = \prod_{j=1}^{n} \mathcal{N}(V_j|0, \sigma^2_V I)
\end{align*}
\]

User-Item Rating Matrix \( R \)

<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>
</tr>
</thead>
<tbody>
<tr>
<td>( u_1 )</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( u_2 )</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( u_3 )</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( u_4 )</td>
<td></td>
<td></td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>( u_5 )</td>
<td></td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Posterior distributions of \( U \) and \( V \) based only on observed ratings

\[ p(U, V|R, \sigma^2_V, \sigma^2_U, \sigma^2_R) \]
User-Tag Tagging Matrix Factorization

Conditional over the observed tagging data

\[ p(C|U, T, \sigma_C^2) = \prod_{i=1}^{m} \prod_{k=1}^{o} [\mathcal{N}(c_{ik}|g(U_i^T T_k), \sigma_C^2)]_C^{i k} \]

- U: user latent feature matrix,
- T: tag latent feature matrix.
- \( U_i^T T_k \): predicted value of the model.

<table>
<thead>
<tr>
<th></th>
<th>t_1</th>
<th>t_2</th>
<th>t_3</th>
<th>t_4</th>
<th>t_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>u_1</td>
<td>4</td>
<td>32</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u_2</td>
<td></td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u_3</td>
<td>3</td>
<td>33</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>u_4</td>
<td></td>
<td></td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>u_5</td>
<td>5</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

User-Tag Tagging Matrix C

Posterior distributions of U and T

\[ p(U, T|C, \sigma_U^2, \sigma_T^2, \sigma_C^2) \]

Jack:
- action (20), animation (20),
- romantic (1)
Item-Tag Tagging Matrix Factorization

<table>
<thead>
<tr>
<th></th>
<th>t₁</th>
<th>t₂</th>
<th>t₃</th>
<th>t₄</th>
<th>t₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>i₁</td>
<td>14</td>
<td>20</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i₂</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>i₃</td>
<td>13</td>
<td>23</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i₄</td>
<td></td>
<td>13</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>i₅</td>
<td>15</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Titanic: romance (20), bittersweet (20), action (1)

Item-Tag Tagging Matrix D

Posterior distributions of V and T

\[ p(V, T | D, \sigma_D^2, \sigma_T^2, \sigma_V^2) \]
# TagRec Framework

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>User latent feature matrix</td>
</tr>
<tr>
<td>V</td>
<td>Item latent feature matrix</td>
</tr>
<tr>
<td>T</td>
<td>Tag latent feature matrix</td>
</tr>
<tr>
<td>R</td>
<td>User-item rating matrix</td>
</tr>
<tr>
<td>C</td>
<td>User-tag tagging matrix</td>
</tr>
<tr>
<td>D</td>
<td>Item-tag tagging matrix</td>
</tr>
</tbody>
</table>

![Diagram of TagRec Framework]
Experimental Analysis

• MovieLens 10M/100K data set:
  – Provided by GroupLens research
  – Online movie recommender service MovieLens (http://movielens.umn.edu)

• Statistics:
  – Ratings: 10,000,054
  – Tags: 95,580
  – Movies: 10,681
  – Users: 71,567
Experimental Analysis

- MAE comparison with other approaches (a smaller MAE means better performance)

<table>
<thead>
<tr>
<th>Training Data</th>
<th>UMEAN</th>
<th>IMEAN</th>
<th>Dimensionality = 10</th>
<th>Dimensionality = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UMEAN</td>
<td>IMEAN</td>
<td>SVD</td>
<td>PMF</td>
</tr>
<tr>
<td>80%</td>
<td>0.7686</td>
<td>0.7379</td>
<td>0.6169</td>
<td>0.6162</td>
</tr>
<tr>
<td>50%</td>
<td>0.7710</td>
<td>0.7389</td>
<td>0.6376</td>
<td>0.6354</td>
</tr>
<tr>
<td>30%</td>
<td>0.7742</td>
<td>0.7399</td>
<td>0.6617</td>
<td>0.6599</td>
</tr>
<tr>
<td>20%</td>
<td>0.7803</td>
<td>0.7416</td>
<td>0.6813</td>
<td>0.6811</td>
</tr>
<tr>
<td>10%</td>
<td>0.8234</td>
<td>0.7484</td>
<td>0.7315</td>
<td>0.7127</td>
</tr>
</tbody>
</table>

UMEAN: mean of the user’s ratings
IMEAN: mean of the item’s ratings
SVD: A well-known method in Netflix competition
PMF: Salakhutdinov and Mnih in NIPS’08
Experimental Analysis

• RMSE comparison with other approaches (a smaller RMSE value means a better performance)

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Baseline Methods</th>
<th>Dimensionality = 10</th>
<th>Dimensionality = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UMEAN</td>
<td>IMEAN</td>
<td>SVD</td>
</tr>
<tr>
<td>80%</td>
<td>0.9779</td>
<td>0.9440</td>
<td>0.8087</td>
</tr>
<tr>
<td>50%</td>
<td>0.9816</td>
<td>0.9463</td>
<td>0.8330</td>
</tr>
<tr>
<td>30%</td>
<td>0.9869</td>
<td>0.9505</td>
<td>0.8636</td>
</tr>
<tr>
<td>20%</td>
<td>1.0008</td>
<td>0.9569</td>
<td>0.8900</td>
</tr>
<tr>
<td>10%</td>
<td>1.1587</td>
<td>0.9851</td>
<td>0.9703</td>
</tr>
</tbody>
</table>
Contribution of Chapter 3

- Propose a factor analysis approach, referred to as TagRec, by utilizing both users’ rating information and tagging information based on probabilistic matrix factorization
- Overcome the data sparsity problem and non-flexibility problem confronted by traditional collaborative filtering algorithms
Introduction

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Structure of Thesis

- Participant
  - User
    - Item
      - Chap. 3
    - User
      - Chap. 4
  - Item
    - Chap. 5
  - Consumption Goods
    - Item
      - Characteristic
      - Chap. 6
Problem and Motivation

• Social Tagging System
Problem and Motivation

- Tagging:
  - Judgments on resources
  - Users’ personal interests
Problem and Motivation

• Providing an **automatic interest-based user recommendation service**
Challenge

• How to model users’ interests?
• How to perform interest-based user recommendation?

Diagram:
- Alex
- Bob
- Grey
- Tom

Connections:
- Alex to Bob
- Grey to Tom
UserRec: User Interest Modeling

• Triplet: user, tag, resource

<table>
<thead>
<tr>
<th>URL</th>
<th><a href="http://www.nba.com">http://www.nba.com</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Tags of user 1</td>
<td>Basketball, nba</td>
</tr>
<tr>
<td>Tags of user 2</td>
<td>Sports, basketball, nba</td>
</tr>
</tbody>
</table>

• Observations of tagging activities:
  - **Frequently** used user tags can be utilized to characterize and capture users’ interests
  - If **two tags** are used by one user to annotate one URL at the same time, it is very likely that these **two tags** are **semantically related**
UserRec: User Interest Modeling

• User Interest Modeling:
  – Generate a weighted tag-graph for each user
  – Employ a community discovery algorithm in each tag-graph
UserRec: User Interest Modeling
UserRec: User Interest Modeling

- Generate a weighted tag-graph for each user:

<table>
<thead>
<tr>
<th>URL</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://espn.go.com">http://espn.go.com</a></td>
<td>basketball, nba, sports</td>
</tr>
<tr>
<td><a href="http://msn.foxsports.com">http://msn.foxsports.com</a></td>
<td>basketball, nba, sports</td>
</tr>
<tr>
<td><a href="http://www.ticketmaster.com">http://www.ticketmaster.com</a></td>
<td>sports, music</td>
</tr>
<tr>
<td><a href="http://freemusicarchive.org">http://freemusicarchive.org</a></td>
<td>music, jazz, blues</td>
</tr>
<tr>
<td><a href="http://www.wwoz.org">http://www.wwoz.org</a></td>
<td>music, jazz, blues</td>
</tr>
</tbody>
</table>

![Tag-graph](image)
UserRec: User Interest Modeling

- Employ community discovery in tag-graph
  - Optimize modularity
  - If the fraction of within-community edges is no different from what we would expect for the randomized network, then modularity will be zero
  - Nonzero values represent deviations from randomness

```
<table>
<thead>
<tr>
<th>Tag</th>
<th>NBA</th>
<th>Sports</th>
<th>Music</th>
<th>Jazz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ball</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
```

Tag-graph

```
<table>
<thead>
<tr>
<th>Tag</th>
<th>NBA</th>
<th>Sports</th>
<th>Music</th>
<th>Jazz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ball</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
```

Two communities
**Interest-based User Recommendation**

- Representing topics of user with a random variable
  - Each **community** discovered is considered as a **topic**
  - A **topic** consists of several **tags**
  - Importance of a topic is measured by the **sum of number of times each tag is used** in this topic
  - Employ **maximum likelihood estimation** to calculate the probability value of each topic of a user

- A Kullback-Leibler divergence (**KL-divergence**) based method to calculate the **similarity** between **two users** based on their topics’ probability distributions
Experimental Analysis

• **Data Set:**
  – Delicious

• **Statistics:**

<table>
<thead>
<tr>
<th>Users</th>
<th>Bookmarks</th>
<th>Network*</th>
<th>Fans**</th>
</tr>
</thead>
<tbody>
<tr>
<td>366,827</td>
<td>49,692,497</td>
<td>425,069</td>
<td>395,415</td>
</tr>
</tbody>
</table>

* This is the total number of users in all users’ personal networks.
** This is the total number of fans of all users.
Experimental Analysis

• **Memory-based** collaborative filtering methods:
  – Person correlation coefficient (PCC)
  – PCC-based similarity calculation method with significance weighting

• **Model-based** collaborative filtering methods:
  – Probabilistic matrix factorization
  – Singular value decomposition
  – After deriving the latent feature matrices, we still need to use memory-based approaches on derived latent feature matrices: SVD-PCC, SVD-PCCW, PMF-PCC, PMF-PCCW
## Experimental Analysis

Comparison with approaches those are based on URLs *(a larger value means a better performance for each metric)*

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Memory-Based Approaches</th>
<th>Model-Based Approaches</th>
<th>UserRec</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCC</td>
<td>PCCW</td>
<td>SVD-PCC</td>
</tr>
<tr>
<td>Precision@R</td>
<td>0.0717</td>
<td>0.1490</td>
<td>0.0886</td>
</tr>
<tr>
<td>MAP</td>
<td>0.1049</td>
<td>0.1874</td>
<td>0.1218</td>
</tr>
<tr>
<td>Bpref</td>
<td>0.0465</td>
<td>0.1148</td>
<td>0.0568</td>
</tr>
<tr>
<td>MMVRR</td>
<td>0.0626</td>
<td>0.1154</td>
<td>0.0710</td>
</tr>
</tbody>
</table>

Comparison with approaches those are based on Tags *(a larger value means a better performance for each metric)*

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Memory-Based Approaches</th>
<th>Model-Based Approaches</th>
<th>UserRec</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCC</td>
<td>PCCW</td>
<td>SVD-PCC</td>
</tr>
<tr>
<td>Precision@R</td>
<td>0.1495</td>
<td>0.3168</td>
<td>0.1540</td>
</tr>
<tr>
<td>MAP</td>
<td>0.1816</td>
<td>0.3444</td>
<td>0.1898</td>
</tr>
<tr>
<td>Bpref</td>
<td>0.1132</td>
<td>0.2395</td>
<td>0.1170</td>
</tr>
<tr>
<td>MMVRR</td>
<td>0.1129</td>
<td>0.1943</td>
<td>0.1151</td>
</tr>
</tbody>
</table>
Contribution of Chapter 4

• Propose the User Recommendation (UserRec) framework for user interest modeling and interest-based user recommendation

• Provide users with an automatic and effective way to discover other users with common interests in social tagging systems
Learning with Social Media

Introduction

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Structure of Thesis

- Participant
  - User
    - Item
      - Chap. 3
    - User
      - Chap. 4
  - Item
    - Chap. 5
  - Characteristic
    - Chap. 6

Social Media

Consumption Goods
Problem and Motivation

• Social media systems with Q&A functionalities have accumulated large archives of questions and answers
  – Online Forums
  – Community-based Q&A services
Problem and Motivation

Query:
Q1: How is Orange Beach in Alabama?

Question Search:
Q2: Any ideas about Orange Beach in Alabama?

Question Suggestion:
Q3: Is the water pretty clear this time of year on Orange Beach?
Q4: Do they have chair and umbrella rentals on Orange Beach?

Topic: travel in orange beach
Results of Our Model

• Why can people only use the air phones when flying on commercial airlines, i.e. no cell phones etc.?

• Results of our model:

1. Why are you supposed to keep cell phone off during flight in commercial airlines? (Semantically equivalent)

2. Why don’t cell phones from the ground at or near airports cause interference in the communications of aircraft? (Semantically related)

3. Cell phones and pagers really dangerous to avionics? (Semantically related)

Interference of aircraft
Problem and Motivation

• Benefits
  – Explore information needs from different aspects
    • “Travel”: beach, water, chair, umbrella
  – Increase page views
    • Enticing users’ clicks on suggested questions
  – Relevance feedback mechanism
    • Mining users’ click through logs on suggested questions
Challenge

- Traditional **bag-of-words approaches** suffer from the shortcoming that they **could not bridge the lexical chasm** between **semantically related questions**
Document Representation

• Document representation
  – Bag-of-words
    • Independent
    • Fine-grained representation
    • Lexically similar
  – Topic model
    • Assign a set of latent topic distributions to each word
    • Capturing important relationships between words
    • Coarse-grained representation
    • Semantically related
TopicTRLM in Online Forum

- TopicTRLM – Topic-enhanced Translation-based Language Model
Learning with Social Media

**TopicTRLM in Online Forum**

\[
P(q \mid D) = \prod_{w \in q} P(w \mid D)
\]

TRLM score: BoW

\[
P(w \mid D) = \gamma P_{trlm}(w \mid D) + (1 - \gamma) P_{lda}(w \mid D)
\]

LDA score: topic model

- **q**: a query, **D**: a candidate question
- **w**: a word in query
- **\( \gamma \)**: parameter balance weights of BoW and topic model
- **Jelinek-Mercer smoothing**
**TopicTRLM in Online Forum**

- **TRLM**
  \[
P_{\text{trlm}}(w \mid D) = \frac{|D|}{|D| + \lambda} P_{\text{mx}}(w \mid D) + \frac{\lambda}{|D| + \lambda} P_{\text{mle}}(w \mid C)
  \]
  \[
P_{\text{mx}}(w \mid D) = \beta P_{\text{mle}}(w \mid D) + (1 - \beta) \sum_{t \in D} T(w \mid t) P_{\text{mle}}(t \mid D)
  \]
  - C: question corpus, \(\lambda\): Dirichlet smoothing parameter
  - T(w|t): word to word translation probabilities

- **Use of LDA**
  \[
P_{\text{lda}}(w \mid D) = \sum_{z=1}^{K} P(w \mid z) P(z \mid D)
  \]
  - K: number of topics, z: a topic
TopicTRLM in Online Forum

• Estimate $T(w|t)$
  – IBM model 1, monolingual parallel corpus
  – Questions are focus of forum discussions, questions posted by a thread starter (TS) during the discussion are very likely to explore different aspects of a topic

• Build parallel corpus
  – Extract questions posted by TS, question pool $Q$
  – Question-question pairs, enumerating combinations in $Q$
  – Aggregating all $q$-$q$ pairs from each forum thread
TopicTRLM-A in Community-based Q&A

• Best answer for each resolved question in community-based Q&A services is always readily available
• Best answer of a question could also explain the semantic meaning of the question
• Propose TopicTRLM-A to incorporate answer information
TopicTRLM-A in Community-based Q&A
Experiments in Online Forum

• Data set
  – Crawled from TripAdvisor
  – TST_LABEL: labeled data for 268 questions
  – TST_UNLABEL: 10,000 threads at least 2 questions posted by thread starters
  – TRAIN_SET: 1,976,522 questions, 971,859 threads
    • Parallel corpus to learn T(w|t)
    • LDA training data
    • Question repository
Experiments in Online Forum

• Performance comparison (a larger value in metric means better performance)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>LDA</th>
<th>QL</th>
<th>TR</th>
<th>TRLM</th>
<th>TopicTRLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P@R$</td>
<td>0.2411</td>
<td>0.3370</td>
<td>0.4135</td>
<td>0.4555</td>
<td>0.5140</td>
</tr>
<tr>
<td>MAP</td>
<td>0.3684</td>
<td>0.4089</td>
<td>0.4629</td>
<td>0.5029</td>
<td>0.5885</td>
</tr>
<tr>
<td>MRR</td>
<td>0.5103</td>
<td>0.5277</td>
<td>0.5311</td>
<td>0.5317</td>
<td>0.5710</td>
</tr>
</tbody>
</table>

• LDA performs the worst, coarse-grained
• TRLM > TR > QL
• TopicTRLM outperforms other approaches
Experiments in Community-based Q&A

• **Date Set**
  - Yahoo! Answers
  - “travel” category
  - “computers & internet” category
Experiments in Community-based Q&A

Performance of different models on category “computers & internet”
(a larger metric value means a better performance)

<table>
<thead>
<tr>
<th>Methods</th>
<th>MAP</th>
<th>Bpref</th>
<th>MRR</th>
<th>P@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.2397</td>
<td>0.136</td>
<td>0.2767</td>
<td>0.1594</td>
</tr>
<tr>
<td>QL</td>
<td>0.346</td>
<td>0.2261</td>
<td>0.416</td>
<td>0.2594</td>
</tr>
<tr>
<td>TRLM</td>
<td>0.3532</td>
<td>0.2368</td>
<td>0.4271</td>
<td>0.2777</td>
</tr>
<tr>
<td>TopicTRLM</td>
<td>0.4235</td>
<td>0.2755</td>
<td>0.5559</td>
<td>0.3197</td>
</tr>
<tr>
<td>TopicTRLM-A</td>
<td>0.6228</td>
<td>0.4673</td>
<td>0.7745</td>
<td>0.5467</td>
</tr>
</tbody>
</table>
Contribution of Chapter 5

• Propose question suggestion, which targets at suggesting questions that are semantically related to a queried question
• Propose the TopicTRLM which fuses both the lexical and latent semantic knowledge in online forums
• Propose the TopicTRLM-A to incorporate answer information in community-based Q&A
Learning with Social Media

Introduction

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Structure of Thesis

- Participant
- Social Media
- Consumption Goods

User
- Item
  - Chap. 3
- User
  - Chap. 4
- Item
  - Chap. 5
- Characteristic
  - Chap. 6
Challenge of Question Analysis

- Questions are ill-phrased, vague and complex
  - Light-weight features are needed
- Lack of labeled data
Problem and Motivation

• “Web-scale learning is to use available large-scale data rather than hoping for annotated data that isn’t available.”

-- Alon Halevy, Peter Norvig and Fernando Pereira
Problem and Motivation

Social Signal

- commenting
- rating
- voting
- Community wisdom

Knowledge

Learning with Social Media
Problem and Motivation

• Whether we can utilize social signals to collect training data for question analysis with NO manual labeling
• Question Subjectivity Identification (QSI)
• Subjective Question
  – One or more subjective answers
  – What was your favorite novel that you read?
• Objective Question
  – Authoritative answer, common knowledge or universal truth
  – What makes the color blue?
Social Signal

- **Like**: like an answer if they find the answer useful
- **Subjective**
  - Answers are opinions, different tastes
  - Best answer receives similar number of likes with other answers
- **Objective**
  - Like an answer which explains universal truth in the most detail
  - Best answer receives higher likes than other answers
Social Signal

• Vote: users could vote for best answer
• Subjective
  – Vote for different answers, support different opinions
  – Low percentage of votes on best answer
• Objective
  – Easy to identify answer contains the most fact
  – Percentage of votes of best answer is high
Social Signal

- **Source:** references to authoritative resources
  - Only available for objective question that has fact answer
- **Poll and Survey**
  - User intent is to seek opinions
  - Very likely to be subjective
Social Signal

• Answer Number: the number of posted answers to each question varies
  • Subjective
    – Post opinions even they notice there are other answers
  • Objective
    – May not post answers to questions that have received other answers since an expected answer is usually fixed
• A large answer number indicates subjectivity
• HOWEVER, a small answer number may be due to many reasons, such as objectivity, small page views
Feature

- Word
- Word n-gram
- Question Length
- Request Word
- Subjectivity Clue
- Punctuation Density
- Grammatical Modifier
- Entity
Experiments

• Dataset
  – Yahoo! Answers, 4,375,429 questions with associated social signals
  – Ground truth: adapted from Li, Liu and Agichtein 2008
CoCQA utilizes some amount of unlabeled data, but it could only utilize a small amount (3,000 questions).

Effectiveness of collecting training data using well-designed social signals.

These social signals could be found in almost all CQA.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>0.6596</td>
</tr>
<tr>
<td>CoCQA</td>
<td>0.6861 (+4.20%)</td>
</tr>
<tr>
<td>L + V + PS + AN + S</td>
<td>0.6626 (+0.45%)</td>
</tr>
<tr>
<td>L</td>
<td>0.5714 (−13.37%)</td>
</tr>
<tr>
<td>V + PS + AN + S</td>
<td>0.6981 (+5.84%)</td>
</tr>
<tr>
<td>PS + AN + S</td>
<td>0.6915 (+4.84%)</td>
</tr>
<tr>
<td>V + PS + AN</td>
<td>0.7214 (+9.37%)</td>
</tr>
<tr>
<td>V + AN</td>
<td>0.7201 (+9.17%)</td>
</tr>
<tr>
<td>AN + S</td>
<td>0.7038 (+6.70%)</td>
</tr>
</tbody>
</table>
Experiments

Better performance using word n-gram compared with word
Social signals achieve on average 12.27% relative gain
Experiments

<table>
<thead>
<tr>
<th>Precision</th>
<th>ngram</th>
<th>ngram + qlength</th>
<th>ngram + rword</th>
<th>ngram + syclue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6596</td>
<td>0.6896</td>
<td>0.6834</td>
<td>0.6799</td>
</tr>
<tr>
<td>ngram</td>
<td>0.7000</td>
<td>0.6950</td>
<td>0.6801</td>
<td>0.6995</td>
</tr>
<tr>
<td>+ pdensity</td>
<td>+ gmodifie</td>
<td>+ entity</td>
<td>heuristic</td>
<td>+ heuristic</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>features</td>
<td></td>
</tr>
</tbody>
</table>

Adding any heuristic feature to word n-gram improve precision

Combining heuristic feature and word n-gram achieves 11.23% relative gain over n-gram
Contribution of Chapter 6

- Propose an approach to collect training data automatically by utilizing social signals in community-based Q&A sites without involving any manual labeling.
- Propose several light-weight features for question subjectivity identification.
Learning with Social Media

Introduction

Background

Item Recommendation with Tagging Ensemble

User Recommendation via Interest Modeling

Item Suggestion with Semantic Analysis

Item Modeling via Data-Driven Approach

Conclusion and Future Work
Conclusion

- Modeling **users’ interests** with respect to their behavior, and **recommending items or users** they may be interested in
  - TagRec
  - UserRec

- Understanding **items’ characteristics**, and grouping items that are **semantically related** for better addressing users’ information needs
  - Question Suggestion
  - Question Subjectivity Identification
Future Work

• TagRec
  – Mine explicit relations to infer some implicit relations

• UserRec
  – Develop a framework to handle the tag ambiguity problem

• Question Suggestion
  – Diversity the suggested questions

• Question Subjectivity Identification
  – Sophisticated features: semantic analysis
Publications: Conferences (7)


Publications: Journals (2), Under Review (1)

• Journals

• Under Review
• Thanks!
• Q & A
FAQ

• FAQ: Chapter 3
• FAQ: Chapter 4
• FAQ: Chapter 5
• FAQ: Chapter 6
FAQ: Chapter 3

- An example of a recommender system
- MAE and RMSE equations
- Parameter sensitivity
- Tag or social network
- Intuition of maximize the log function of the posterior distribution in Eq. 3.10 of thesis

Back to FAQ
An Example of A Recommender System

Have some personal preferences.

Get some recommendations.

These recommendations are based on items you own and more.

1. **Pattern Recognition and Machine Learning (Information Science and Statistics)**
   by Christopher M. Bishop (Oct 1, 2007)
   Average Customer Review: ★★★★★ (22)
   In Stock
   List Price: $95.00
   Price: $68.81
   $2 used & new from $54.97
   [Add to Cart] [Add to Wish List]
   I own it  Not interested  Rate this item
   Recommended because you rated Pattern Classification (2nd Edition) and more

   by Trevor Hastie (Jun 6, 2009)
   Average Customer Review: ★★★★★ (22)
   In Stock
   List Price: $89.86
   Price: $71.96
   $2 used & new from $68.32
   [Add to Cart] [Add to Wish List]
   I own it  Not interested  Rate this item
   Recommended because you rated Pattern Classification (2nd Edition) and more

3. **Computer Manual in MATLAB to Accompany Pattern Classification, Second Edition**
   by David G. Stork (April 8, 2004)
   Average Customer Review: ★★★★★ (6)
   In Stock
   List Price: $46.65
   Price: $40.99
   $2 used & new from $27.00
   [Add to Cart] [Add to Wish List]
   I own it  Not interested  Rate this item
   Recommended because you rated Pattern Classification (2nd Edition)
MAE and RMSE

• Mean absolute error (MAE)

\[ MAE = \frac{\sum_{ij} |r_{ij} - \hat{r}_{ij}|}{N} \]

• Root mean squared error (RMSE)

\[ RMSE = \sqrt{\frac{\sum_{i,j} (r_{i,j} - \hat{r}_{i,j})^2}{N}} \]
Parameter Sensitivity

![Graphs showing parameter sensitivity](image-url)
Tag or Social Network?

• What is the difference of incorporating tag information and social network information?
• Answer: both tagging and social networking could be considered as user behavior besides rating. They explain users’ preferences from different angles. The proposed TagRec framework could not only incorporate tag information, but also could utilize social network information in a similar framework.
Intuition of maximize the log function of the posterior distribution in Eq. 3.10 of thesis

- The idea of **maximize the log function of the posterior distributions** is equivalent to **maximize the posterior distributions** directly, because the **logarithm** is a **continuous strictly increasing function** over the range of the likelihood. The reason why I would like to maximize the posterior distributions is that after Bayesian inference, I need to calculate the conditional distributions to get the posterior distributions, e.g.: \( p(R|U,V) \), \( R \) is the observed ratings, and \( U, V \) are parameters. To estimate the \( U, V \), I use the **maximum likelihood estimation** to estimate the parameter space, thus I need to **maximize the conditional distributions** \( P(R|U,V) \). So this is the reason why I have to maximize the log function in my approach.
FAQ: Chapter 4

- What is modularity?
- Comparison on Precision@N
- Comparison on Top-K accuracy
- Comparison on Top-K recall
- Distribution of number of users in network
- Distribution of number of fans of a user
- Relationship between # fans and # bookmarks
- Why we use the graph mining algorithm instead of some simple algorithms, e.g. frequent mining

Back to FAQ
What is Modularity?

• The concept of modularity of a network is widely recognized as a good measure for the strength of the community structure

\[
Q = \frac{1}{2m} \sum_{i,j} \left[ A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)
\]

\[
k_i = \sum_k A_{ik} \quad m = \frac{1}{2} \sum_{i,j} A_{ij}
\]

\(A_{ij}\) is the weight between node \(i\) and node \(j\)

\(\delta(c_i, c_j)\) is 1 if node \(i\) and node \(j\) belong to the same community
Comparison on Precision@N

![Graph showing comparison of Precision@N for different methods.](image)

- UserRec
- PCCW@URL
- SVD-PCCW@URL
- PMF-PCCW@URL
- PCCW@Tag
- SVD-PCCW@Tag
- PMF-PCCW@Tag

Back to FAQ
Comparison on Top-K Accuracy

Top-K Accuracy of UserRec and Other Approaches

Percentage of Users Add At Least One Top-K

Top-K Recommended Users

UserRec, PCCW@URL, SVD-PCCW@URL, PMF-PCCW@URL, PCCW@Tag, SVD-PCCW@Tag, PMF-PCCW@Tag

Back to FAQ
Comparison on Top-K Recall

Top-K Recall of UserRec and Other Approaches

Percentage of Users Covered by Top-K

Top-K Recommended Users
Distribution of Number of Users in Network
Distribution of Number of Fans of A User
Relationship Between # Fans, # bookmarks

Relation between number of bookmarks and number of fans

Number of bookmarks

Number of fans
Why we use the graph mining algorithm instead of some simple algorithms, e.g. frequent itemset mining

- We use community discovery algorithm on each tag-graph, and could accurately capture users’ interests on different topics. The algorithm is efficient, and the complexity is $O(n\log^2 n)$. While frequent itemset mining is suitable for mining small itemset, e.g., 1, 2, 3 items in each set. However, each topic could contain many tags.
FAQ: Chapter 5

- Experiments on word translation
- Dirichlet smoothing
- Build monolingual parallel corpus in community-based Q&A
- An example from Yahoo! Answers
- Formulations of TopicTRLM-A
- Data Analysis in online forums
- Performance on Yahoo! Answers “travel”
Experiments on Word Translation

- **Word translation**

<table>
<thead>
<tr>
<th>Words</th>
<th>IBM 1</th>
<th>LDA</th>
<th>IBM 1</th>
<th>LDA</th>
<th>IBM 1</th>
<th>LDA</th>
<th>IBM 1</th>
<th>LDA</th>
<th>IBM 1</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>shore</td>
<td>shore</td>
<td>park</td>
<td>park</td>
<td>condo</td>
<td>condo</td>
<td>beach</td>
<td>beach</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>beach</td>
<td>groceri</td>
<td>drive</td>
<td>hotel</td>
<td>beach</td>
<td>south</td>
<td>resort</td>
<td>slope</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>snorkel</td>
<td>thrift</td>
<td>car</td>
<td>stay</td>
<td>area</td>
<td>north</td>
<td>what</td>
<td>jet</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>island</td>
<td>supermarket</td>
<td>how</td>
<td>time</td>
<td>unit</td>
<td>shore</td>
<td>hotel</td>
<td>snowboard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>kauai</td>
<td>store</td>
<td>area</td>
<td>area</td>
<td>island</td>
<td>pacif</td>
<td>water</td>
<td>beaver</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>condo</td>
<td>nappi</td>
<td>where</td>
<td>recommend</td>
<td>maui</td>
<td>windward</td>
<td>walk</td>
<td>huski</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>area</td>
<td>tesco</td>
<td>walk</td>
<td>beach</td>
<td>rent</td>
<td>seaport</td>
<td>area</td>
<td>steamboat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>water</td>
<td>soriana</td>
<td>time</td>
<td>nation</td>
<td>owner</td>
<td>alabama</td>
<td>room</td>
<td>jetski</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>boat</td>
<td>drugstore</td>
<td>ride</td>
<td>tour</td>
<td>shore</td>
<td>opposit</td>
<td>snorkel</td>
<td>powder</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>ocean</td>
<td>mega</td>
<td>hotel</td>
<td>central</td>
<td>rental</td>
<td>manor</td>
<td>restaur</td>
<td>hotel</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **IBM 1**: semantic relationships of words from semantically related questions
- **LDA**: co-occurrence relations in a question
Dirichlet Smoothing

• Bayesian smoothing using Dirichlet priors
  – A language model is a multinomial distribution, for which the conjugate prior for Bayesian analysis is the Dirichlet distribution
  – Choose the parameters of the Dirichlet to be
    \[(\mu p(w_1 \mid C), \mu p(w_2 \mid C), \ldots, \mu p(w_n \mid C))\]
  – Then the model is given by
    \[
p_\mu(w \mid d) = \frac{c(w; d) + \mu p(w \mid C)}{\sum_{w' \in V} c(w'; d) + \mu}
    \]
Build Monolingual Parallel Corpus in Community-based Q&A

• Aggregate question title and question detail as a monolingual parallel corpus
An Example from Yahoo! Answers

Resolved Question

Should we buy brandable domains?
I personally don't really invest in brandable domain names. What you guys suggest: Is it worth to buy brandable domains?
4 hours ago

Best Answer - Chosen by Asker
Yeah you should buy them. If its comes in your budget, you should go for them and I guess I'm familiar with a website which will let you to have the domains at reasonable prices, https://www.email.biz/ !!
3 hours ago

Asker's Rating: *****
thanks for the help i really need it.

Best answer available

Back to FAQ
TopicTRLM-A in Community-based Q&A

\[
P(q|(Q, A)) = \prod_{w \in q} P(w|(Q, A)),
\]

\[
P(w|(Q, A)) = \epsilon P_{trlm}(w|(Q, A)) + (1 - \epsilon) P_{lda}(w|Q)
\]

Lexical score \hspace{2cm} Latent semantic score
TopicTRLM-A in Community-based Q&A

\[
P_{trlm}(w|(Q, A)) = \frac{|(Q, A)|}{|(Q, A)| + \lambda} P_{mx}(w|(Q, A)) + \frac{\lambda}{|(Q, A)| + \lambda} P_{mle}(w|C),
\]

\[
P_{mx}(w|(Q, A)) = \eta P_{mle}(w|Q) + \theta \sum_{t \in Q} T(w|t) P_{mle}(t|Q) + \mu P_{mle}(w|A)
\]

Dirichlet smoothing

Question LM score
Question translation model score
Answer ensemble

Back to FAQ
Data Analysis in Online Forums

- **Data Analysis**
- **Post level**

<table>
<thead>
<tr>
<th># Threads</th>
<th># Threads that have replied posts from TS</th>
<th>Average # replied posts from TS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,412,141</td>
<td>566,256</td>
<td>1.9</td>
</tr>
</tbody>
</table>

- **Forum discussions are quite interactive**
- **Power law**

![Distribution of replied posts from thread starter](image.png)

Back to FAQ
Performance on Yahoo! Answers “travel”

Performance of different models on category “travel” (a larger metric value means a better performance)

<table>
<thead>
<tr>
<th>Methods</th>
<th>MAP</th>
<th>Bpref</th>
<th>MRR</th>
<th>P@R</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.1345</td>
<td>0.0612</td>
<td>0.1616</td>
<td>0.0675</td>
</tr>
<tr>
<td>QL</td>
<td>0.316</td>
<td>0.1902</td>
<td>0.388</td>
<td>0.2048</td>
</tr>
<tr>
<td>TRLM</td>
<td>0.3222</td>
<td>0.2034</td>
<td>0.3923</td>
<td>0.2234</td>
</tr>
<tr>
<td>TopicTRLM</td>
<td>0.3615</td>
<td>0.244</td>
<td>0.4406</td>
<td>0.2644</td>
</tr>
<tr>
<td>TopicTRLM-A</td>
<td>0.467</td>
<td>0.3167</td>
<td>0.5963</td>
<td>0.387</td>
</tr>
</tbody>
</table>
FAQ: Chapter 6

- **Examples of subjective, objective questions**
- **Benefits of performing question subjectivity identification**
- **How to define subjective and object questions**

Back to FAQ
Examples of Subjective, Objective Questions

• Question subjectivity identification

• Subjective
  – What was your favorite novel that you read?
  – What are the ways to calm myself when flying?

• Objective
  – When and how did Tom Thompson die? He is one of the group of Seven.
  – What makes the color blue?
Benefits of Performing QSI

- More accurately identify similar questions
- Better rank or filter the answers
- Crucial component of inferring user intent
- Subjective question --> Route to users
- Objective question --> Trigger AFQA
How to define subjective and object questions

- Ground truth data was created using Amazon’s Mechanical Turk service. Each question was judged by 5 qualified Mechanical Turk workers. Subjectivity was decided using majority voting.
- Linguistic people are good at manual labeling.
- Compute science people should focus on how to use existing data to identify subjective/objective questions, such as social signals, answers, etc. Not focus on manual labeling.