Scalable Mining of Social Data using Stochastic Gradient Fisher Scoring

Jeon-Hyung Kang and Kristina Lerman
USC Information Sciences Institute
Information Overload in Social Media

2,500,000,000,000,000,000 (2.5 quintillion) bytes of data is created everyday! -IBM estimates 2012

27% of total U.S. internet time is spent on social networking sites -Experian 2012

Social Networking is the No. 1 Online Activity in the U.S. (based on the Average time U.S. consumers spent with digital media per day) -Mashable 2013

Ruff, J. (2002). Information Overload: Causes, Symptoms and Solutions. Harvard Graduate School of Education’s Learning Innovations Laboratory (LILA)
Recommendations in Social Media

- Social Media provides a solution by allowing users to subscribe/follow updates from specific users.
- But Information Overload will comeback when users follows many others.

Filtering & Recommendation!

<table>
<thead>
<tr>
<th>Flexibility (Model)</th>
<th>Scalability (Computation)</th>
<th>Sparsity (Data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variety of Data</td>
<td>2.5 quintillion</td>
<td>Long tail</td>
</tr>
<tr>
<td>Easily Extensible</td>
<td>Information Overload</td>
<td>Distribution</td>
</tr>
</tbody>
</table>

Social Media

User

Recommendations
## User-Item Rating Prediction Problem

<table>
<thead>
<tr>
<th>Users</th>
<th>Ratings</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daniel</td>
<td>4.0</td>
<td><img src="image" alt="Breaking Bad" /> <img src="image" alt="Walking Dead" /> <img src="image" alt="The Avengers" /> <img src="image" alt="Thor" /> <img src="image" alt="Three Will Turn West" /> <img src="image" alt="Much Ado" /> <img src="image" alt="Love Actually" /> <img src="image" alt="Hugo" /></td>
</tr>
<tr>
<td>Sara</td>
<td>5.0</td>
<td><img src="image" alt="Breaking Bad" /> <img src="image" alt="Walking Dead" /> <img src="image" alt="The Avengers" /> <img src="image" alt="Thor" /> <img src="image" alt="Three Will Turn West" /> <img src="image" alt="Much Ado" /> <img src="image" alt="Love Actually" /> <img src="image" alt="Hugo" /></td>
</tr>
<tr>
<td>Bob</td>
<td>1.0</td>
<td><img src="image" alt="Breaking Bad" /> <img src="image" alt="Walking Dead" /> <img src="image" alt="The Avengers" /> <img src="image" alt="Thor" /> <img src="image" alt="Three Will Turn West" /> <img src="image" alt="Much Ado" /> <img src="image" alt="Love Actually" /> <img src="image" alt="Hugo" /></td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td><img src="image" alt="Breaking Bad" /> <img src="image" alt="Walking Dead" /> <img src="image" alt="The Avengers" /> <img src="image" alt="Thor" /> <img src="image" alt="Three Will Turn West" /> <img src="image" alt="Much Ado" /> <img src="image" alt="Love Actually" /> <img src="image" alt="Hugo" /></td>
</tr>
<tr>
<td>Users</td>
<td></td>
<td><img src="image" alt="Breaking Bad" /> <img src="image" alt="Walking Dead" /> <img src="image" alt="The Avengers" /> <img src="image" alt="Thor" /> <img src="image" alt="Three Will Turn West" /> <img src="image" alt="Much Ado" /> <img src="image" alt="Love Actually" /> <img src="image" alt="Hugo" /></td>
</tr>
</tbody>
</table>

- **Users**: Daniel, Sara, Bob
- **Items**: ![Breaking Bad](image), ![Walking Dead](image), ![The Avengers](image), ![Thor](image), ![Three Will Turn West](image), ![Much Ado](image), ![Love Actually](image), ![Hugo](image)
Probabilistic Matrix Factorization (PMF)

“PMF is a probabilistic linear model with Gaussian observation noise that handles very large and possibly sparse data.”
**Probabilistic Matrix Factorization (PMF)**

“PMF is a probabilistic linear model with Gaussian observation noise that handles very large and possibly sparse data.”

[Salakhutdinov & Mnih 08]
Additionally Available Data

- User
  - Social Network
- Profile
- User
  - Ratings
- Item
- User
  - Source Of Adoptions
- Item
  - Descriptions
- Term
Given the parameters, computing the full posterior of latent variables conditioned on a set of observations is intractable!

⇒ Approximate Posterior Inference
Approximate Posterior Inference

Gibbs Sampling

• Simplest Markov chain Monte Carlo (MCMC) algorithm.
• Sampling each one from its distribution conditioned on the current assignment of all other variables.
• It requires computations over the **whole data set**

Stochastic Gradient Fisher Scoring

• Stochastic Gradient: To approximate the gradient over the whole data set using **min-batch of data**
  - The average of gradient of the log likelihood w.r.t. given mini-batches
• Bayesian Central Limit Theorem (Le Cam, 1986): under suitable regularity conditions, \( P(\theta \mid X) \) approximately equals \( \mathcal{N}(\theta_0, I_N^{-1}) \) as \( N \) becomes very large!

Social Data: characterized by a long-tailed distribution.

The long tail of some distribution having a large number of occurrences far from the “head” or central part of the distribution. The distribution could involve popularities, random numbers of occurrences of events with various probabilities. (i.e. online business, mass media, micro-finance, social network, economic models, and marketing)

The Pareto principle (also known as the 80-20 rule, the law of the vital few, and the principle of factor scarcity) states that, for many events, roughly 80% of the effects come from 20% of the causes.
Why do we care about a long-tail?

Gained popularity in recent times as describing the retailing strategy of selling a large number of unique items in addition to selling fewer popular items.

The distribution and inventory costs of businesses successfully applying this strategy allow them to realize significant profit out of selling small volumes of hard-to-find items to many customers instead of only selling large volumes of a reduced number of popular items.

<table>
<thead>
<tr>
<th>Source of data set</th>
<th>MovieLens 2000</th>
<th>KDD Cup 2012</th>
<th>CiteULike 2013</th>
<th>Twitter 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td># of users</td>
<td>6K users</td>
<td>2.3 mil users</td>
<td>5K users</td>
<td>9.5 mil users</td>
</tr>
<tr>
<td># of items</td>
<td>3.9K movies</td>
<td>6K items</td>
<td>31K items</td>
<td>114K items</td>
</tr>
<tr>
<td># of user-item adoptions</td>
<td>1 million anonymous ratings</td>
<td>5.2 mil adoptions</td>
<td>303K adoptions</td>
<td>13 mil adoptions</td>
</tr>
<tr>
<td>Sparseness</td>
<td>76.6%</td>
<td>98.75%</td>
<td>99.81%</td>
<td>99.87%</td>
</tr>
<tr>
<td>Mean # of adoptions/user</td>
<td>166 items</td>
<td>77 items</td>
<td>9 items</td>
<td>8 items</td>
</tr>
<tr>
<td>Median # of adoptions/user</td>
<td>96 items</td>
<td>58 items</td>
<td>8 items</td>
<td>6 items</td>
</tr>
<tr>
<td>notes</td>
<td>We used the same training and test data set split as in the original dataset.(0.9 mil training and 0.1 mil test)</td>
<td>User profiles (including user keywords, age, gender, tags, number of tweets), item categories (pre-assigned topic hierarchy), item keywords, and user activity.</td>
<td>The public information about users, papers, tags that have been used by users to describe papers, the group information that link users to group by subscription.</td>
<td>707 mil social network links. Collected whole chains of tweets containing URLs to monitor information propagations over the social net. Focused on URLs containing at least 5 terms resulting in 14K users, 6K items, &amp; 121K adoptions.</td>
</tr>
</tbody>
</table>
CPU time per epoch is different with the number of different size of mini batch.

Except the burn-in period, SGFS outperformed Gibbs Sampling.
SGFS Result on Social Media Data

KDD Cup 2012

CiteULike

Twitter

Likelihood

Recall@N
Hybrid-Stochastic Gradient Fisher Scoring

- \( N_{\text{CLT}} \): the number of observations for the Bayesian Central Limit Theorem to hold

- Hidden variables with enough observations will be inferred using SGFS, while those with few observations will be inferred using GD.

- Can handle data that follows a long-tail distribution.
SGFS Result on Social Media Data

Scalability of distributed SGFS

Recall@N
Conclusions

- Explored **SGFS** for mining social data and showed that algorithm scales up with both the number of processors and the size of mini-batch.
  - **Failed** to learn good models of *long-tailed distributed data*, resulting in poor prediction performance.
- Many social media data sets have long tailed distributions, containing sparse regions with few observations per variable.
- Proposed **hSGFS** algorithm by combining SGFS and Gradient descent to provide more efficient performance.
  - Showed significant **performance improvement** in both **speed** and **prediction accuracy**.
  - We believe that **hSGFS** is an initial step to further work on efficient MCMC sampling based on stochastic gradients on long-tailed distributed data.
Related Work

- **Stochastic Inference**
  - Bayesian learning via stochastic gradient langevin dynamics. [Welling and The 11]
  - Bayesian posterior sampling via stochastic gradient fisher scoring. [Ahn et al 12]
  - Stochastic Variational Inference. [Hoffman et al 12]

- **Collaborative filtering methods for item recommendations**
  - Probabilistic Matrix Factorization [Salakhutdinov & Mnith 08]
  - Recommend new items that were liked by similar users [Ma et al. 08, Chua et al 12]
  - Improve explanatory power of recommendations by extending LDA & SN matrix factorization [Wang and Blei 11, Prushotham et al 12]
  - Limited attention in social media [Kang & Lerman 13]