Key Lessons Learned Building Recommender Systems For Large-scale Social Networks
The world’s largest professional network

Over 50% of members are now international

165M+ *

New members *

34th
Most visit website worldwide (Comscore) *

>2M
Company pages **

75%
Fortune 500 Companies use LinkedIn to hire **

*as of Nov 4, 2011
**as of June 30, 2011
LinkedIn Homepage

Powered by Recommendations!
LinkedIn Recruiting Solutions

Reach out to people who match your job description for: Director of Product Management

Daniel Tunkelang
Principal Data Scientist at LinkedIn
San Francisco Bay Area • Computer Science
InMails (1) - Notes (2) - Projects (2) - Views (12)
Sort by: Connections •
Current: Principal Data Scientist at LinkedIn; Member at New York CTO Club
Past: Tech Lead at Google; Chief Scientist at Endeca; Consultant at Context Integration; Research Intern / Consultant at IBM T. J. Watson Research Center; Research Intern at AT&T Bell Labs
Education: Carnegie Mellon University; Massachusetts Institute of Technology; St. Hilda's & St. Hugh's
Network Info: 5 recommendations, 500+ connections

James G. (Jimi) Shanahan
Experienced Data Mining and Machine Learning Consultant
San Francisco Bay Area • Information Technology and Services
InMails (9) - Projects (6) - Views (11)
Current: Search and Advertising Consultant at Dip; Statistical Machine Learning Consultant at Statistical Machine Learning Consultant
Past: Executive Director of Research in Local and Mobile Search at AT&T Interactive; Web search ranking consultant at Searchme; Technical Advisor at Kouns Ltd; Chief Scientist at Turn Inc.; Principal Research Scientist at Calixvance Corporation; Research Scientist at Xerox Research Corporation Europe; Senior Researcher at Laboratory for International Fuzzy Engineering; Senior Researcher at Laboratory for Fuzzy Engineering, A Software Engineer at Mitsubishi
Education: University of Bristol; University of Limerick
Network Info: 1 recommendation, 500+ connections

Michael Wu PhD
Principal Scientist at Lithium Technologies
San Francisco Bay Area • Computer Science
InMails (1) - Projects (4) - Views (2)
Current: Principal Scientist at Lithium Technologies
Past: Research Scholar in Gallant Lab at UC Berkeley; US Dept. of Energy CSOF Researcher at Los Alamos National Laboratory; Admin Assistant in Fore Lab at UC Berkeley; Laboratory Assistant in Owen Lab at UC Berkeley; Programmer Analyst in Math Department at UC Berkeley; Lecturer and Study Group Leader at The Student Learning Center at UC Berkeley;
The Recommendations Opportunity

Jobs You May Be Interested In

Talent Match

CAP

Similar Profiles

Companies
Recommendations, similar companies search, peer companies, and company browse maps, company products and services browse maps

Related search

Profile browse maps

Connections

Events You May Be Interested In

Network updates

Jobs browse maps

News

Groups
Recommendations, similar groups search

Similar jobs

Referral Engine

Similar Profiles

News
LinkedIn Today: "Top Headlines for You"

LinkedIn

Similar Profiles

Events You May Be Interested In

Network updates

Network updates

Referral Engine
Outline

• Some Fundamentals of RecSys
• Recommendation Lifecycle
• Right Time Recommendations at Scale
What is a Recommender System?

A Recommender selects a product that if acquired by the “buyer” maximizes value of both “buyer” and “seller” at a given point in time

…Must make a strategic sense!
Ingredients of a Recommender System

A Recommender processes information and transforms it into actionable knowledge.

Data: 25%
Algorithms: 5%
Business Logic and Analytics: 20%
User Experience: 50%

Strands, RecSys’09
User Experience Matters

1. Understand user intent + interest
   \[\rightarrow\] Define, model, leverage \[\leftarrow\] I-2
2. Be in the user flow
3. Location, location, location
4. Set right expectations (“You May….”)
5. Explain recommendations
6. Interact with the user
### User Intent, User Flow, Location, Message

<table>
<thead>
<tr>
<th>Product</th>
<th>Intent and flow</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jobs browsemaps</td>
<td>Job description page: <em>job seeker</em></td>
<td>7X application rate</td>
</tr>
<tr>
<td></td>
<td>After having applied to a job: <em>job applicant</em></td>
<td></td>
</tr>
<tr>
<td>Jobs You May be Interested In</td>
<td>Homepage: <em>all</em></td>
<td>10X application rate</td>
</tr>
<tr>
<td></td>
<td>Job homepage: <em>job seeker</em></td>
<td></td>
</tr>
<tr>
<td>Jobs You May be Interested In</td>
<td>Job homepage center</td>
<td>5X application rate</td>
</tr>
<tr>
<td></td>
<td>Job homepage right rail</td>
<td></td>
</tr>
</tbody>
</table>

**Item based collaborative filtering:**  
→ Followers

**Content based:**  
→ Leaders
User Intent Modeling

1. Intent labeling
   - Job seeker, recruiter, outbound professional, sales, content consumer, content creator, networker, brander

2. Build normalized propensity model for each intent

3. Job Seeker:
   - Active, receptive, non-seeking
   - Score = Ordered Probit(behavior, content, and network)
   - 10X application rate between active and passive receptive
Lifecycle of a Recommendation

1. Define *clear* objective → Metric to optimize
2. Build, train and evaluate new model
   • Generate labeled and feature data
3. A/B test (and sometimes optimize) online
4. Rollout to all users and track if successful
5. Go back to 2!
Recommendation Objectives – Beyond CTR

Suggest relevant groups that one is more likely to participate in

Suggest skilled candidates who will likely respond to hiring managers inquiries
MOO – Multiple Objective Optimization

• Pareto-based multi-objective machine learning framework with two-layer scoring approach:
  1. Optimize for relevance only
  2. Then optimize for secondary objective while maintaining relevance

→ TalentMatch+Flightmeter
= 55% lift in inmail response rate
Propensity Score – The Group Case

• Social learning: people are followers
  “People learn new behavior through observational learning of the social factors in their environment. If people observe positive, desired outcomes in the observed behavior, then they are more likely to model, imitate, and adopt the behavior themselves.”

• Facebook Study ’09
  “Newcomers may not recognize the value of contribution. [...] We find support for social learning: newcomers who see their friends contributing go on to share more content themselves.

  *Burke et al: “Feed Me: Motivating Newcomer Contribution in Social Network Sites,” CHI’09*
Propensity Score – The Group Case

- One idea borrowed from expertise networks in online communities

- Group Activity = Directed Graph(V,A)
  - V = group members
  - i → j if j comments on i’s initial posting

Zhang et al., “Expertise Networks in Online Communities: Structure and Algorithms,” WWW’07
Propensity Score – The Group Case

• **Indegree**: Prestige
  - Engaging more is more prestigious

• **Outdegree**: Influence
  - Postings generating engagement

• Both follow a Power Law distributions

\[
P(d_{in}) \sim d_{in}^{-\alpha} \quad P(d_{out}) \sim d_{out}^{-\beta}
\]

• \(\Rightarrow\) Group propensity score = \(F(\alpha, \beta)\)
Platform

• Scaling innovation
  • Cross-leverage improvements between products
  • Shared knowledgebase

• Maintainability
  • Production serving and tracking
  • Infrastructure for complete upgrades

• Performance
  • Billions of sub second computations
LinkedIn Recommendation Engine

Recommendation Entities
- TalentMatch
- People Browse Map
- Similar Profiles
- Referral Center
- Jobs You May be interested in
- Jobs Browse Map
- Similar Jobs
- GYML
- Groups Browse Map
- Similar Groups

Recommendation Types
- Shared
- Dynamic
- Unified
- Core Service

Content & Behavior Analysis
- (R-T) Feature Extraction, Entity Resolution & Enrichment

Collaborative Filtering
- (R-T) matching computations

Popularity

User Feedback
- Ads
- Companies Searches
- News Events
- ... and more to come

Products
- People
- Jobs
- Groups

API
Quick Overview of the Recommendation Process

**Input Entities**
Jobs, people, ads, news, events,…

**Target Entities**
People, jobs, groups,…

**Extracted Features**

**Matching**

**Filtering Business Rules**
**Matching**

**Binary**
- Exact match
- Exact match in bucket

**Soft Match**
- $v_1 = tf \times idf$
- $\cos \Theta = \frac{v_1 \times v_2}{|v_1| \times |v_2|}$

**Comparison Table**

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialty → Specialty</td>
<td>0.58</td>
</tr>
<tr>
<td>Skills → Skills</td>
<td>0.94</td>
</tr>
<tr>
<td>Title → Title</td>
<td>0.26</td>
</tr>
<tr>
<td>Seniority → Seniority</td>
<td>0.18</td>
</tr>
<tr>
<td>Summary → Summary</td>
<td>0.98</td>
</tr>
<tr>
<td>Title → Related Title</td>
<td>0.16</td>
</tr>
<tr>
<td>Education → Education</td>
<td>0.40</td>
</tr>
</tbody>
</table>

**Example**

- **Christian Posse**
  - Principal Data Scientist and Product Manager at LinkedIn
  - San Francisco Bay Area | Information Technology and Services
  - **Title**: Data Scientist
  - **Specialty**: Data Science
  - **Education**: Computer Science
  - **Experience**: 10 years
  - **Location**: San Francisco, CA
  - **Industry**: Technology
  - **Skills**: Data Analysis, Machine Learning

- **Daniel Tunkelang**
  - Principal Data Scientist at LinkedIn
  - Mountain View, CA | Computer Software
  - **Title**: Data Scientist
  - **Specialty**: Data Science
  - **Education**: Computer Science
  - **Experience**: 10 years
  - **Location**: Mountain View, CA
  - **Industry**: Technology
  - **Skills**: Data Analysis, Machine Learning
Importance weight vector (Skills-> Skills)

Similarity score vector (Skills-> Skills)

Normalization, Scoring & Ranking

\[
\frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i})}}
\]

Filtering
Location
Company
Industry

Feedback
Technologies
Data Sources for Feature Engineering

Social Graphs

Content

http://inmaps.linkedinlabs.com/

Behavior

PVs, Queries, Actions (clicks)
Feature Engineering – Entity Resolution

• Companies
  ‘IBM’ has **8000+** variations
  - ibm – ireland
  - ibm research
  - T J Watson Labs
  - International Bus. Machines

• Huge impact on the business and UE
  • Ad targeting
  • TalentMatch
  • Referrals
Feature Engineering - Enrichment

- Zip code mapped to Regions
- How sticky are those locations?
- Huge impact on the business and UE
  - Job Seeker, Recruiter

*Most vs Least sticky regions*
*what percentage of people stay in the same region when switching companies*
Feature Engineering - Enrichment

• How sticky are those locations?
  • Region similarity based on profiles or network
  • Region transition probability

→ predict individuals propensity to migrate and most likely migration target

• Impact on job recommendations
  • 20% lift in views/viewers/applications/applicants
Feature Engineering - Enrichment

• Time-Based Seniority
• (Inferred) Skills
• User intent/need
  • Job seeker (active, passive, not)
• Related entities (title, company, industry)
• Virtual Profiles for communities
Data Labeling

- **Historical data**
  - Similar objective
  - Unrelated processes, e.g., same session search selection
    - reduce presentation bias, position bias
    - What about intent bias?

- **Random suggestions**
  - Great with ads, company follows
  - Not for products with high cost of bad recommendations
    - jobs, alumni groups, …
  - Not for similar recommendations

- **Crowdsourced manual labeling**
  - Very challenging
  - Pairwise comparison more suited than absolute rating → higher costs
  - Expertsourced manual labeling
A/B Testing
Is Option A Better Than Option B? Let’s Test

Beware of
- novelty effect
- cannibalization
- potential biases (time, targeted population)
- random sampling destroying the network effect

Enjoy testing furiously!

Don’t forget to A/A test first

("Seven Pitfalls to Avoid when Running Controlled Experiments on the Web", KDD’09
“Framework and Algorithms for Network Bucket Testing” WWW’12 submission)