Diversity in Recommender Systems
Week 2: The Problems

Toni Mikkola, Andy Valjakka, Heng Gui, Wilson Poon
Review

- diversification happens by “searching from further away”
- balancing diversity and relevance
- diversity from user behaviour and ratings
  - no need for semantic information
- Yu et al. 2009
  - notion of explanations
    - set of items from which the recommendation comes
- Boim et al. 2011
  - Priority cover trees (PCT) and priority medoids
    - structures that denote the distance
Common Drawbacks

- Cold start
- “Harry Potter” problem
- Yu et al. 2009: need for manual threshold tweaking
- Need for semantic information in order to consider diversity
- Boim et al. 2011: relying heavily on item ratings
- Some algorithms a bit complex to be practical:
  - Yu 2009: Greedy
  - Boim 2011: Priority Medoids being NP Hard + PCT population
Cold start

Cold Start means, the new users can’t get good recommendation or new items can’t be recommended because of lacking related informations.

E.g. The user can’t get any recommendation when he first enter the del.icio.us.

The new item will not be recommend to the user because it has no relevance.

Cold start problem make the recommender system don’t have enough items to recommend, even worse in diversification.
Some website try to solve this problem by giving questionnaire to new users. At least, they can get some informations for recommending. Amazon will give their new product to the reviewer (active user) to get the initial rating and comments.
Harry Potter Problem

With a item-based collaborative filtering approach, such as Amazon's "People who bought this also bought", the system runs the risk of recommending Harry Potter to everyone just because most people have bought the Harry Potter book.

This kind of very popular items, will always score highly in our algorithm, so it will often be recommended no matter who is buying whatever items.

This recommendation is useless because “Harry Potter” items don’t have a real relevance or connection with the based item. It will makes the system less efficient and perform worse.

E.g recommending a plastic bag to customer in supermarket.
Yu et al: Threshold tweaking

All the algorithm in this paper heavily rely on the given threshold. The threshold means how many items should be recommended because of “diversity”.

But the paper did not illustrate what threshold is good.

Actually, best threshold will be different in different contexts like movie website, restaurant website or music apps.

We may need more research or experiment on this topic. E.g. what threshold in music app recommender system can make users satisfy most.
Common Drawbacks

● Cold start
● “Harry Potter” problem
● Yu et al. 2009: need for manual threshold tweaking
● Need for semantic information in order to consider diversity
● Boim et al. 2011: relying heavily on item ratings
● Some algorithms a bit complex to be practical:
  ○ Yu 2009: Greedy
  ○ Boim 2011: Priority Medoids being NP Hard + PCT population
Semantics for diversity: interrelated work

Sequel Diversity: # of distinct series in item set. \(\text{sequelDiversity}(I_k) = \frac{|\{\text{series}(i) | i \in I_k\}|}{k}\)

- Without semantics, how do we know, e.g. how movies are interrelated?
  - Movies in the same universe? e.g. Batman The Dark Knight and Batman Vs Superman are not technically sequels but in the same universe
  - Spiritual sequels? Robocop and Starship troopers
  - Not a sequel but title is misleading: Jackie Chan’s Drunken Master I, II and III

- Well, we don’t, because algorithms don’t care!
  - Assumes relation is clustered with relevant user / item groups & data structures

- However, semantics are still needed to **evaluate** diversity.
Boim et al: Item Ratings

- Certain people, not all, give ratings
  - Thus, a large population is not reflected
  - May consider to limit data to ratings only from critics, magazines, review websites

- Subjective to user (groups)
  - Niches according to tastes: Local sports, children’s books
  - Critics have a more demanding palate than normal users
  - Bias, e.g. Apple (iPhone) fans in love with the company and never use other products

- Numerically quantifying one’s rating
  - How often are perfect scores given?
  - Will scoring be re-adjusted once users become more experienced?

- Some ratings are given tongue-in-cheek
  - e.g. five stars for The Room
Common Drawbacks

- Cold start
- “Harry Potter” problem
- Need for semantic information in order to consider diversity
- Yu et al. 2009: need for manual threshold tweaking
- Boim et al. 2011: relying heavily on item ratings

- Some algorithms a bit complex to be practical:
  - Boim 2011: Priority Medoids being NP Hard + PCT population
  - Yu 2009: Greedy
Boim et al: Medoid & PCT computability

- Priority medoids are NP hard
  - Meaning it’s at least as hard to compute as e.g. TSP
- Priority cover tree
  - Regular cover tree can just be populated arbitrarily
  - PCT requires *ordered insertion*
    - ...which requires sorting which isn’t talked about
  - Also *tight insertion*
    - necessary only for specific purpose
Boim et al. 2011 - Results

- Are the algorithms significantly better?
  - "PCT-R is best in sequel diversity" according to paper
- Swap algorithm required manual threshold tuning
  - What threshold to use
- To their credit, PCT-R seems to be the best overall
  - User study details omitted
- What about different rating systems?
Implicit and explicit feedback

- **Explicit feedback**
  - Hard to collect, i.e. users don’t rate or “like” so many items -> Sparsity
  - Easier to recognize items of which the user likes
  - How to collect opinions?
    - If rating scale is wide, users’ different rating behaviour could affect the recommendations
    - If rating scale is narrow, users can’t express enough well their emotions about the items

- **Implicit feedback**
  - Easier to collect, i.e. if users don’t rate quite often, at least they click, view or buy an item
  - Harder to recognize items of which users really like. The purpose of the users’ action can change and is not clear. E.g. the user buys an item for someone else.
  - Data can be noisy depending on which actions the rec. system collects
Sparsity

- Growing challenge in CF-based recommender system when the number of users or items grows
  - Less users or items benefit from recommendations, even less users or items benefit from diversity in recommendations.
  - Sparse matrix needs effective computation and explanations don’t come free, they need extra computation after candidate items for recommendations have been generated.

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>del.icio.us friendship avg. rec./user</td>
<td>2.2</td>
<td>8.7</td>
<td>9.8</td>
</tr>
<tr>
<td></td>
<td>avg. explanation/rec.</td>
<td>1</td>
<td>1.2</td>
</tr>
<tr>
<td>del.icio.us shared-url avg. rec./user</td>
<td>4.89</td>
<td>9.41</td>
<td>9.93</td>
</tr>
<tr>
<td></td>
<td>avg. explanation/rec.</td>
<td>1</td>
<td>1.71</td>
</tr>
<tr>
<td>Y! Movies shared-movie avg. rec./user</td>
<td>5.61</td>
<td>9.72</td>
<td>9.95</td>
</tr>
<tr>
<td></td>
<td>avg. explanation/rec.</td>
<td>1.56</td>
<td>3.74</td>
</tr>
</tbody>
</table>

Table 5: Statistics of recommendations and explanations with number of recommendations capped at 10.

Figure 2: Average costs for generating recommendations and explanations.
Evaluation of Diversification in rec. systems

- Offline evaluation methods don’t work well with diversity
  - WHY OFFLINE
    - Offline methods are cheap and repeatable, statistical, scalable
  - WHY NOT OFFLINE
    - Offline methods are not the “real thing”, no user feedback, hard to measure quality of diversity, novelty and serendipity
  - Optimal top-K recommendations with diversity don’t exist
  - Standard Precision/Recall can’t measure benefits of diversity in recommendations.
    - Recall: percentage of relevant items guessed
      \[
      \frac{\#(\text{Hidden} \cap \text{Recommended})}{\# \text{ Hidden}}
      \]
    - Precision: average quality of each recommendation
      \[
      \frac{\#(\text{Hidden} \cap \text{Recommended})}{\# \text{ Recommended}}
      \]
Online Evaluation and User Studies

- E.g. A/B tests, feedback queries
- WHY ONLINE EVALUATION
  - Easier to measure the quality of diversity, user interaction, real thing - real time
- WHY NOT ONLINE EVALUATION
  - Expensive, time consuming, interferes with business, not repeatable, continuous monitoring
- User studies’ results can be unclear, not general and hard to implement
Conclusion

- Cold start
- Lack of semantics
- Inconclusiveness of results
- Insufficient technicalities
  - Next time we may consider the omitted details
Thank you!