SCENE
A Scalable Two-Stage Personalized News Recommendation System

Lei Li, Dingding Wang, Tao li, Daniel Knox, Balaji Padmanabhan

School of Computing and Information Sciences
Florida International University
& DailyMe Inc.
SCENE: A SCalable Two-Stage PErsontalized News REcommendation System

SIGIR’11 @Beijing
Personalized News Recommendation

SCENE: A SCalable Two-Stage PErsontalized News REcommendation System

SIGIR'11 @Beijing
Why SCENE?

• *How to effectively select news articles for recommendation?*
  - Consider online users’ navigation behaviors
  - Explore unique characteristics of news articles, e.g., short shelf lives, value of immediacy

• *How to effectively present the recommended new articles?*
  - Facilitate users’ navigation and exploration
  - Extensible to multi-dimensional recommendation with flexible reading operations, such as drill-down and roll-up

• *How to efficiently handle large scale news collection?*
  - Reduce computational redundancy as much as possible

• *How to construct high-quality user profiles?*
  - Utilize the encapsulation of different yet related information
SCENE: A Scalable Two-Stage Personalized News Recommendation System

Submodularity Modeling

Locality Sensitive Hashing + Hierarchical Clustering

Two-level News Hierarchy

News Content Similar Access Patterns Preferred Named Entities

SIGIR’11 @Beijing
Prior Approaches

• **Content-based Filtering**
  - Compare news articles with user’s reading histories
  - Easy to implement and explain
  - Not enough to capture user’s interest
    - Newsjunkie (Gabrilovich et al., *WWW*, 2004)

• **Collaborative Filtering**
  - Use a group of readers’ past rating behaviors (1 – read, 0 – not read) to predict rating on newly-published articles
  - Easy to capture users’ reading behaviors
  - Suffer from cold-start problem
    - GoogleNews (Das et al., *WWW*, 2007)

• **Hybrid Approaches**
  - Combine collaborative filtering and content-based filtering
    - YahooNews (Chu and Park, *WWW*, 2009)
Our Contribution – SCENE

- **A principled framework for news selection**
  - Consider the interestingness of news articles with respect to users
  - Model news selection as a budgeted maximum coverage problem
  - Balance the novelty and diversity of recommendation result

- **A novel two-level representation**
  - Present users a two-level news hierarchy
    - First Level: General topic categories
    - Second Level: Representative news articles
  - Facilitate users’ navigation experience

- **Multi-factor high-quality user profiling**
  - Incorporate three different yet related information (news content, similar access patterns and preferred named entities)
  - Maximally capture the exact reading interests of users
SCENE – System Overview

**Clustering on news articles**

- Divide news collection into small groups using LSH purely based on news content;
- Hierarchical clustering on small news groups, and cut the dendrogram;
- Employ probabilistic language models to summarizing news articles in each intermediate cluster and small news groups within the cluster.
SCENE – System Overview

User Profiling

- Encapsulate three different yet related information
  - News Content (topic distribution)
  - Access Pattern (Click Behavior)
  - Named Entity (Entity Extraction)
SCENE – System Overview

**Personalized News Selection**

- Identify clusters with content similar to a given user’s profile;
- Select the news group in each cluster with content most similar to the user’s profile;
- In each news group, employ submodularity model to select articles.
SCENE – News Articles Clustering

- Clustering over news articles

  ➤ Reduce *computational redundancy* using LSH

  - Decompose news articles to shingles
    - *k*-shingle: a sequence of continuous words
    - choose *k* to guarantee that the probability of any given shingle appearing in any article is low
  - Minhashing on shingle vectors
    - use minhashing technique (Indyk, *SIAM on Discrete Algorithms*, 1999) to generate signatures
    - the probability that two articles have the same signature is equal to the Jaccard similarity between these two
  - Locality sensitive hashing (*LSH*) on minhash signatures
    - use *LSH* (Gionis et al., *VLDB*, 1999) to hash signatures into a big hash table (small news groups)
• Clustering over news groups and topic detection

Facilitate users’ navigation using News Hierarchy

- Hierarchical clustering on small news groups (average-link)
- Dendrogram cut on news hierarchy
- Topic Detection on intermediate clusters and small news groups using LDA
**SCENE – User Profiling**

- *Encapsulation* of three *different yet related* information

  - Content Summary, Similar Access Pattern, Preferred Named Entity

  - Why different?
    - Content summary describes the general interest of users via topic distribution;
    - Similar access pattern includes the click behavior of users;
    - Preferred named entity involves when, where, what happened, who are involved, etc.

  - Why related?
    - The topic distribution learned from the reading history is likely to be related to the list of entities in the profile;
    - Content and named entity might contribute to access patterns that the two users have.
SCENE – User Profiling (cont.)

- How can we get them?

  - Content Summary, Similar Access Pattern, Preferred Named Entity
    - Content summary
      - Summarize a user’s reading history using language models, e.g., LDA.
    - Similar access pattern
      - Compute the similarity between users based on their click behavior (collaborative filtering);
      - Select users with similarity to a given user larger than a threshold, and observe their click behavior.
    - Preferred named entity
      - Extract named entities from the user’s reading history using NLP tool – GATE.
SCENE – Personalized Recommendation

- Rough matching with the user’s interest to Groups

1st level of representation by content matching

- Compute content similarity between the user’s profile and the intermediate clusters’ summary
- Dynamically choose intermediate clusters as the 1st level representation
- Select the news group most similar to the user’s profile in each cluster
• Fine-grained matching with the user’s interest to *Articles*

>$2^{nd}$ level of representation by profile matching

Possible solutions (consider info capsule):

- Greedily select news articles from each cluster in terms of similarity.
  - Information redundancy!
  - Ignored users’ reading behavior!

- Model article selection as a contextual bandit problem ([Li et al., *WWW*, 2010])
  - long-term effect of recommendation.
SCENE – Personalized Recommendation (cont.)

• Fine-grained matching with the user’s interest to Articles

**2\textsuperscript{nd} level of representation by profile matching**

Our solutions (consider info capsule):

- Use submodularity to model the recommendation as a \textit{budgeted maximum coverage} problem;
- Define a \textit{quality function} to evaluate the overall quality of recommendation result;
- Greedily select articles to \textit{maximize} the quality function;

How to select news articles to recommend?
Introduction to Submodularity

★ Submodular Function: Let \( E \) be a finite set and \( f \) be a real valued non-decreasing function defined on the subsets of \( E \) that satisfies

\[
f(T \cup \{\varsigma\}) - f(T) \leq f(S \cup \{\varsigma\}) - f(S)
\]

where \( S \subseteq T \), \( S \) and \( T \) are two subsets of \( E \), and \( \varsigma \in E \setminus T \).

★ Budgeted Maximum Coverage: Given \( E \), where each element is associated with an influence and a cost defined over a domain of these elements and a budget \( B \), the goal is to find out a subset of \( E \) which has the largest possible influence while the total cost does not exceed \( B \).

NP-Hard!!!

(\(1 - 1/e\))-approximation.

(Khuller et al., Information Processing, 1999)
Why Submodularity?

- A topic might involve multiple sub-topics. How to make the recommendation more diverse even for the same topic?
- A user’s interest in a reading session might be regressive over time. How to capture such click behavior?

Explore properties of news set and model users’ click behaviors
• Submodularity modeling for recommendation

**Diversify the result as well as satisfying reader’s need**

**Submodular Atoms**: Given $N$ (the original news group) and $S$ (the selected news set), after selecting a news article into $S$:

- **Representative**
  
  $S$ should be similar to the general topic in $N\backslash S$;

- **Diversity**
  
  The topic diversity should not deviate much in $S$;

- **Satisfaction**
  
  $S$ should further satisfy the user’s reading preference.
SCENE – Personalized Recommendation (cont.)

- Submodularity modeling for recommendation

Diversify the result as well as satisfying reader’s need

Quality Function:

\[
f(S) = \frac{1}{|N \setminus S| \cdot |S|} \sum_{n_1 \in N \setminus S} \sum_{n_2 \in S} \text{sim}(n_1, n_2)
\]

Representative!

\[
+ \frac{1}{\binom{|S|}{2}} \sum_{n_1, n_2 \in S} \text{sim}(n_1, n_2)
\]

Diversity!

\[
+ \frac{1}{|S|} \sum_{n_1 \in S} \text{sim}(u, n_1)
\]

Satisfaction!

Linear combination of submodular functions is still submodular.

(Leskovec, SIGKDD, 2007)
Submodularity modeling for recommendation

- Diversify the result as well as satisfying reader’s need

- Quality Increase:

\[ I(\zeta) = f(S \cup \{\zeta\}) - f(S) \]

- Goal: select a list of news articles which provide the largest possible quality increase within the budget.

Final Recommendation List:

- Select top ranking items within each group, where the number of items selected in one group is proportional to the interest weight of the user on the corresponding topic category.
SCENE – Personalized Recommendation (cont.)

• Ranking adjustment for recommendation list

Consider the properties of news articles (popularity & recency)

★ Popularity: describe how popular the article is, related to # of Visits.

★ Recency: describe how immediate the article is, related to Time.

Adjustment Strategy:

- Integrate the popularity and recency scores for each article in the list;

\[
n_\phi = \frac{n_P - n_{P_{min}}}{n_{P_{max}} - n_{P_{min}}} - \frac{n_I - n_{I_{min}}}{n_{I_{max}} - n_{I_{min}}} \]

- Popularity

- Recency

- Sequentially select two adjacent news items from top to bottom;

- Compare the difference of their dynamic score. If > 0, swap them; otherwise, skip them and continue to compare the next article pair.
SCENE – Experimental Evaluation

• Real-world News Data
  - Gather news and click behaviors from popular news websites for 9 categories [Aug 15th 2010 – Nov 16th, 2010].
  - Preprocessing
    - Remove news articles that are rarely accessed
    - Remove users with infrequent online reading behavior
  - Statistics
    - 112,380 news articles
    - 4,630 online users
    - 1,221 news per day in average

<table>
<thead>
<tr>
<th>category</th>
<th># of Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economy</td>
<td>18,204</td>
</tr>
<tr>
<td>Entertainment</td>
<td>20,460</td>
</tr>
<tr>
<td>Environment</td>
<td>4,147</td>
</tr>
<tr>
<td>Health</td>
<td>5,542</td>
</tr>
<tr>
<td>Law</td>
<td>11,873</td>
</tr>
<tr>
<td>Politics</td>
<td>21,240</td>
</tr>
<tr>
<td>Science</td>
<td>8,815</td>
</tr>
<tr>
<td>Sports</td>
<td>9,562</td>
</tr>
<tr>
<td>World</td>
<td>12,537</td>
</tr>
</tbody>
</table>
• Experimental Setup
  - An offline component for periodically clustering news articles in a time range
    - LSH + Hierarchical Clustering on newly-published articles
  - An online component for dynamically constructing and updating user profiles
    - Encapsulate different resources into users’ profiles
  - An online component for personalized news recommendation
    - Recommend news articles via submodularity modeling
Experimental Methodology

- Select news articles within a time range \([T1,T2]\) for testing
- Regard reading behaviors of \(t < T1\) as users’ reading history
- Randomly choose 100 users from \(t < T1\), and select news articles within \([T1,T2]\) for recommending (top @10, @20, @30)
- Evaluate precision, recall and F-score of the result
## SCENE – Clustering Component Evaluation

- **LSH + Hierarchical Clustering V.S. Direct Clustering**

<table>
<thead>
<tr>
<th>Time Range</th>
<th># of Articles</th>
<th>K-means</th>
<th>Hierarchical (average-link)</th>
<th>“LSH + Hierarchical”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Micro-F1</td>
<td>Macro-F1</td>
<td>Time Cost</td>
</tr>
<tr>
<td>08/15–08/16</td>
<td>2,340</td>
<td>0.4330</td>
<td>0.3744</td>
<td>3min</td>
</tr>
<tr>
<td>08/19–08/22</td>
<td>5,572</td>
<td>0.4365</td>
<td>0.3653</td>
<td>5min</td>
</tr>
<tr>
<td>08/23–08/31</td>
<td>10,985</td>
<td>0.4227</td>
<td>0.3421</td>
<td>10min</td>
</tr>
<tr>
<td>09/01–09/15</td>
<td>18,841</td>
<td>0.3962</td>
<td>0.3302</td>
<td>21min</td>
</tr>
<tr>
<td>09/01–09/30</td>
<td>35,920</td>
<td>0.3783</td>
<td>0.3153</td>
<td>38min</td>
</tr>
<tr>
<td>10/01–11/16</td>
<td>63,659</td>
<td>0.3529</td>
<td>0.2976</td>
<td>59min</td>
</tr>
<tr>
<td>08/15–11/16</td>
<td>112,380</td>
<td>0.3217</td>
<td>0.2711</td>
<td>121min</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>0.3045</td>
<td>0.3280</td>
<td></td>
</tr>
</tbody>
</table>

- **Micro-F1**: weights equally all the articles.
- **Macro-F1**: weights equally all the clusters.

- Efficiency improvement
  - LSH reduces computational redundancy to a great extent.
- Accuracy improvement
  - Represent articles with shingles instead of using words.
SCENE – Profiling Component Evaluation

- Explore combinations of three different yet related resources

- Recommendation purely based on one single aspect cannot effectively capture users’ reading interest;

- Recommendation with named entity involved performs better. (online readers prefer simple but representative named entities.)
Submodularity-based strategy has better performance in terms of precision and recall.

The results of submodularity-based strategy are more stable (with smaller variation).
• Cold-start handling
  - Explore reading behaviors of three different user groups
    Suppose a user reads $N$ news articles per day:
      (i) $N \leq 10$; (ii) $10 < N \leq 50$; (iii) $N > 50$
  - Compare with:
    ✓ Das et al., *WWW*, 2007 (Collaborative Filtering)
    ✓ Liu et al., *IUI*, 2010 (Content Filtering)

  - For the group of $N \leq 10$, SCENE handles the cold-start problem very well.
    ✓ Info capsule of three different yet related resources.
SCENE – News Selection Component Evaluation

- Diversity evaluation

  Given a news set $N$, evaluate the average dissimilarity of $N$:

  $$f_d(N) = \frac{2}{p(p-1)} \sum_{n_i \in N} \sum_{n_j \in N, n_j \neq n_i} (1 - Sim(n_i, n_j))$$

  $p$ – the size of the news set $N$;
  $Sim(n_i, n_j)$ – the news pair similarity of $n_i$ and $n_j$.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Top @10</th>
<th>Top @20</th>
<th>Top @30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Goo</td>
<td>0.4101</td>
<td>0.3074</td>
<td>0.1105</td>
</tr>
<tr>
<td>ClickB</td>
<td>0.4329</td>
<td>0.3128</td>
<td>0.1562</td>
</tr>
<tr>
<td>Bilinear</td>
<td>0.4234</td>
<td>0.2517</td>
<td>0.0933</td>
</tr>
<tr>
<td>Bandit</td>
<td>0.5056</td>
<td>0.4126</td>
<td>0.2925</td>
</tr>
<tr>
<td>SCENE</td>
<td>0.6930</td>
<td>0.6671</td>
<td>0.6059</td>
</tr>
</tbody>
</table>

- SCENE – More diverse!
  
  ✓ Two-stage news recommendation
  
  ✓ Submodularity modeling on news selection
SCENE – Conclusion and Future Work

• **A scalable two-stage personalized news recommender**
  - A principled framework for news selection by considering online readers’ navigation behaviors;
  - A novel two-level representation of recommendation results to facilitate users’ navigation and exploration;
  - A multi-factor user profiling technique to encapsulate three different yet related information of users’ reading history.

• **For future work**
  - Use Map-Reduce framework to extend the scalability of SCENE;
  - Explore users’ interest evolution over time.
Thank you!
Questions?

問題?

Ερώτηση?

Domanda?

문제?

Pregunta?