SCENE : A Scalable Two-Stage Personalized News Recommendation System

Contribute:

Propose a Scalable two-stage Personalize news recommendation approach with a two-level representation

- First level consist various topics relevant to users' preference and divide into group
- 2nd level includes specific news articles and recommend news items

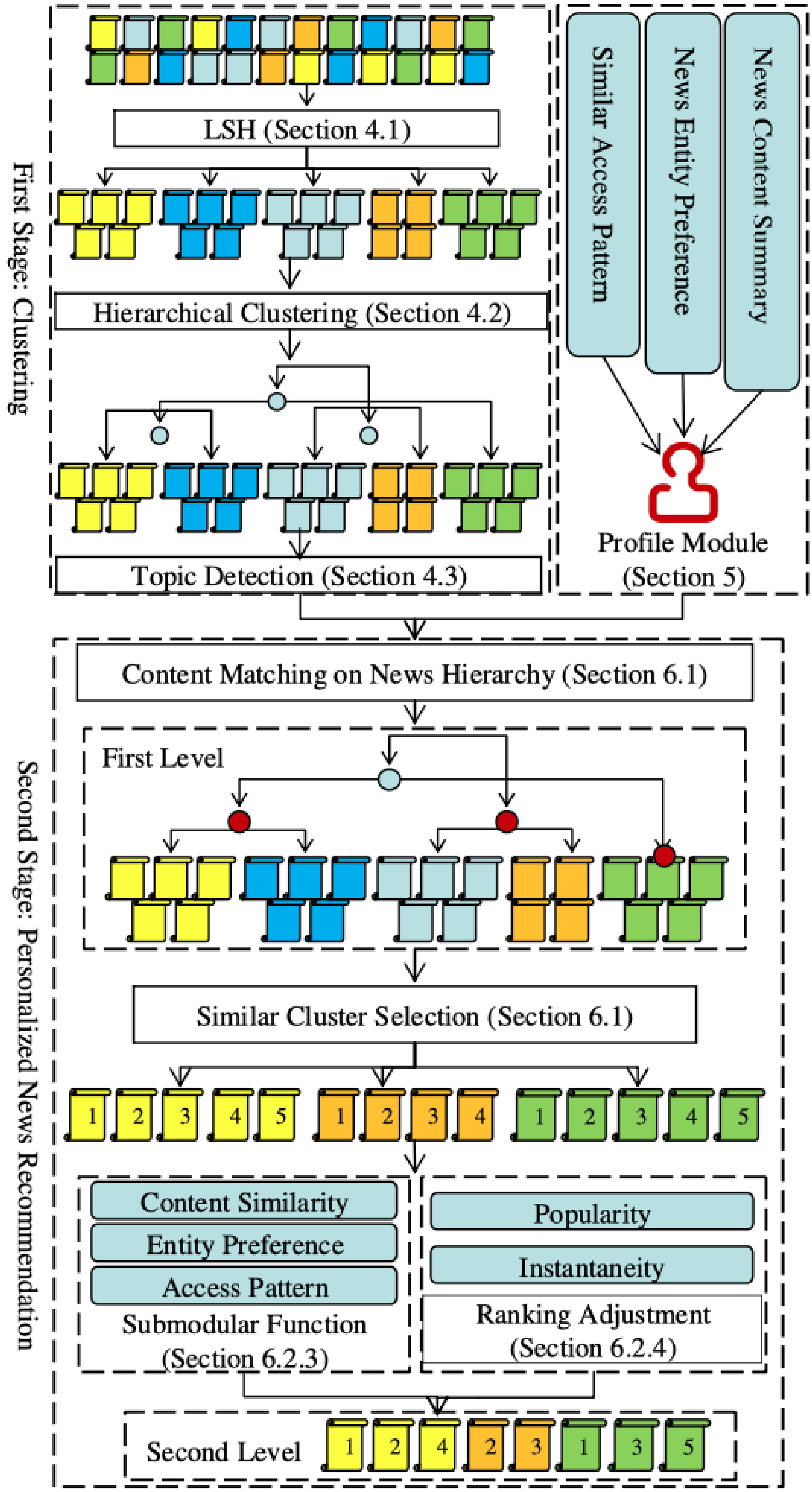
Principled framework for news selection good balance between the novelty and diversity

Multi-factor high-quality user profile construction

SCENE: Consists of 3 Major components:

- Newly-Published News Articles Clustering
  - Partition newly published news into small groups by **Locality Sensitive Hashing**
  - Hierarchically separate groups with average-link. Then LM **Probabilistic Latent Semantic Indexing** (PLSI) and **Latent Dirichlet Allocation** (LDA) applied to summarizing news articles
- User Profile Contruction
  - Constructed in: News topic distribution, Similar access patterns and news entity preference (which are all extracted from user's read history)
- Personalized News Items Recommendation
  - Compare the topic distributions of cluster and news content
  - Then sequentially select the clusters based on Similarity (1^{st} level)
  - Continue compare Similarities of small news group and user's accesed news content (most similar groups as 2^{nd} level)
  - In the most similar groups are model personalized news recommendation by greedy way

Recommendation Framework



Methods in Framework

News article clustering

- Decompose news articles to shingles
  - Preprocessing: remove stop words, tokenize and stemm
  - Shingling articles with $k = 10$ into matrix M with rows is shingles and columns is articles
- MinHasing
  - Contruct Randomized 100 length Minhash signature
- Locality Sensitive Hashing
  - Signatures are initally decomposed into multiple bands, then use standard hash function to hash into a big hash table (band length = 5)
  - Then separated news corpus into small news groups by Jaccard-based

News topic Detection

- Detecting topics of a text corpus done by using probabilistic language models as PLSI or LDA, by extracting a list of represent words from corpus.

User Profile Contruction

- Built by eploration on: News content, Similar access patterns and preferred news entities
  - Each User profile can be present : $\mathcal{U} =< \mathcal{T}, \mathcal{P}, \mathcal{E} >$
    - $\mathcal{T}$ represent topic distribution of news that user access in the past $\{< t_1, w_1 >, < t_2, w_2 >, ...\}$ with corresponding weight
    - $\mathcal{P}$ represent list of user $< u_1, u_2, ... >$ similar access pattens with given user
    - $\mathcal{E}$ represent list of named entities $< e_1, e_2, ... >$ from user's reading history
  - **News content summarize** by use represent users' reading history as the same as the representations for news groups
  - **Access pattern** is built by calulating pairwise similarity between user given and orther user's reading history with Jacard-Sim. Similarity score is predefined by theshold.
  - **Name entities** is built by NLP task. "Each news article is associated with a list of named entities along with their corresponding entity types"

Personalized Recommendation

Interest Matching for Representation Lv.1

- Ranking by cosine similarity between topic distribution of each cluster $\mathcal{T}_\mathcal{C}$ and user's profile $\mathcal{T}_\mathcal{U}$
- Choose the clusters with the score greater than dynamic threshold
- Dig into each cluster and choose the news group most similar to the user's interest

News Selection for Representation Lv.2

- **News Profile** helpful to compare two news, and evaluate how the news item can satisfy the user's reading preference
  - News profile $\mathcal{F}_n =< \mathcal{T}_n, \mathcal{P}_n, \mathcal{E}_n >$ and User's profile $\mathcal{F}_u =< \mathcal{T}_u, \mathcal{P}_u, \mathcal{E}_u >$, and Similarity of them is computed:

$$\mathcal{S}im(\mathcal{F}_n, \mathcal{F}_u) = \frac{\alpha \mathcal{S}im(\mathcal{T}_n, \mathcal{T}_u) + \beta \mathcal{S}im(\mathcal{P}_n, \mathcal{P}_u) + \gamma \mathcal{S}im(\mathcal{E}_n, \mathcal{E}_u)}{\sqrt{\alpha^2 + \beta^2 + \gamma^2}}$$

Where α, β, γ are parameters to control how we trust the corresponding components

$Sim(T_n, T_u)$ is computed by Cosin similarity

$Sim(P_n, P_u), Sim(E_n, E_u)$ is computed by Jaccard similarity

Submodularity

Tyically, a news reader is not interested all of aspects in the given topic. So, news selection strategy can be described as follows: ($\mathcal{N}$ is original news group, $\mathcal{S}$ is selected news set, ζ is the news being selected).

After selecting ζ:

- S should be similar to the general topic in $N \setminus S$;
- The topic topic diversity should not deviate much in S;
- S should provide more satisfaction to the given user's reading preference

Quality function f to evaluate the news set S:

$$f(\mathcal{S}) = \frac{1}{|\mathcal{N} \backslash \mathcal{S}|.|\mathcal{S}|} \sum_{n_1 \in \mathcal{N} \backslash \mathcal{S}} \sum_{n_2 \in \mathcal{S}} sim(n_1, n_2) + \frac{1}{\binom{|S|}{2}} \sum_{\substack{n_1, n_2 \in S \\ n_1 \neq n_2}} -sim(n_1, n_2) + \frac{1}{|\mathcal{S}|} \sum_{n_1 \in \mathcal{S}} sim(u, n_1)$$

n_1 and n_2 denote news items, u denote given user

Three components are involved:

- The first one aims to evaluate the quality of selected news set $\mathcal{S}$ over original
- The second one provides a perspective on how diverse that the top- ics underlying the selected news articles
- The last one is evidence that how much the user's preference is satisfied by $\mathcal{S}$

Ranking Adjustment

The ranking of the selected news articles(popularity and recency) need to be adjusted for more reasonable by normalized two types of properties

Given n, the popularity n_P and the recency n_I can be combined as

$$n_\phi = \frac{n_P - n_{P_{min}}}{n_{p_{max}} - n_{p_{min}}} - \frac{n_I - n_{I_{min}}}{n_{I_{max}} - n_{I_{min}}}.$$

Recency is restricted by time

The generated ranking can emphasize more popular and fresh news items