A Graph-Based Technical Paper Recommender System Using Fused Similarity Metrics - Related Work

By Janelle Blankenburg
Outline

- Introduction
- Related Work
  - Collaborative Filtering Methods
  - Content-Based Filtering Methods
    - Citation-Based Methods
    - Context-Aware Methods
    - Keyword-Based Methods
- Conclusion
- Project Idea
Introduction
Problem Description

- Technical paper recommendation
- General process
  - Identify keywords related to research interests
  - Input keywords into online textual search service
    - Google Scholar, CiteSeer, arXiv
  - Modify keyword list and repeat
Motivation and Challenges

- Simple keyword search is not sufficient
  - Highly technical terms
  - Variations of phrasing of terms
- To alleviate repetitive process, create technical paper recommender systems that rely on more input than simple keywords
Motivation and Challenges Cont.

- Sample Inputs:
  - Authorship details
  - Access data
  - One or multiple papers

- Systems use input data to generate a user model as basis for providing recommendations
Related Work
Collaborative Filtering (CF)

- Provide recommendations to the user based on the preferences of like-minded users.
Content-Based Filtering (CBF)

- Provide recommendations based on similarity of new items to the user’s existing set of items
Citation-Based Methods

These methods use an existing citation graph to make recommendations.
Finding Relevant Papers Based On Citation Relations

Fig. 2. Related papers of $p_1$.

Finding Relevant Papers Based On Citation Relations

- **Local Relation Strength**
  - \( LRS(p_{citer} \rightarrow p_{citee}) = w_r \cdot \frac{|Q_{pcitee}^r|}{|Q_{pcitee}|} \cdot e^{-\frac{d(T_{p_{citer}}, T_{p_{citee}})}{d_M}} \)

- **Global Relation Strength**
  - \( GRS(p_1 \rightarrow p_2) = \sum_{\theta_i \in \theta} [\eta]^{\theta_i} \cdot \sum_{e_{mn} \in \theta_i} \frac{\lambda_{mn}}{|\theta_i|} \)


- Citation Graph
  \[ ((p_i, p_j), (p_j, p_i) \in \text{Edge}\{G_{\tilde{C}}\}) \iff (p_i \in \mathcal{R}_{p_j} \lor p_j \in \mathcal{R}_{p_i}) \]
- Connectivity Matrix
  \[ \tilde{C}_{i,j} = \begin{cases} 
    1 & \text{if } (p_i, p_j) \in \text{Edge}\{G_{\tilde{C}}\} \\
    0 & \text{otherwise} 
  \end{cases} \]
- Correlation Matrix
  \[ C_{i,j} = \frac{\tilde{C}_{i,j}}{\omega_j} \]


- Classic PageRank
  \[ PR(n) = \alpha \cdot \sum_{q: (q,n) \in \mathcal{E}} \frac{PR(q)}{\omega_q} + (1 - \alpha) \cdot \frac{1}{|\mathcal{V}|} \]

- Biased PageRank
  \[ \mathbf{PR} = \alpha \cdot \mathbf{M} \cdot \mathbf{PR} + (1 - \alpha) \cdot \mathbf{d} \]

- PaperRank
  \[ \mathbf{IR} = \alpha \cdot \mathbf{C} \cdot \mathbf{IR} + (1 - \alpha) \cdot \mathbf{d} \]

Automatically Building Research Reading Lists

Automatically Building Research Reading Lists

- Half-life Utility:
  - $R_a = \sum_i u_{a, i} 2^{(i-1)/(\alpha-1)}$
  - $R = \frac{\sum_{a \in T} R_a}{|T|R_{max}}$

Context-Aware Methods

These methods use words around a given citation to recommend which paper to cite in that specified location.
Utilizing Context in Generative Bayesian Models for Linked Corpus

Figure 1: Bayesian Network for (a) linked-LDA, (b) link-PLSA-LDA, (c) cite-LDA and (d) cite-PLSA-LDA

A Supervised Learning Method for Context-Aware Cit. Rec. in a Large Corpus

- $f: (\overrightarrow{d}, \overrightarrow{d}, c*) \rightarrow [0, 1]$
- 30 Features in 4 categories:
  - Popularity-Based
  - Candidate Meta-Data
  - Author-Based
  - Meta-Data Similarity

A Supervised Learning Method for Context-Aware Cit. Rec. in a Large Corpus

- Full ML-based Recommender System
  - Uses all 30 Features
  - Combines LibSVM, Logistics Model Tree, LogitBoost, Artificial Neural Networks using backpropagation, Random Forest, Rotation Forest, and AdaBoost
  - Average estimated probabilities of each classifier

Keyword-Based Methods

These methods analyze various components of a paper, such as the title and abstract, in order to provide recommendations.
PubMed: A Probabilistic Topic Based Model for Content Similarity

- Uses terms from title and abstract
- Fit Poisson distributions for each term in the document:

\[
w_t = \left( 1 + \eta \left( \frac{\mu}{\lambda} \right)^k e^{-(\mu-\lambda)l} \right)^{-1} \sqrt{idf_t}
\]

\[
Sim(c, d) = \sum_{t=1}^{N} w_{t, c} \cdot w_{t, d}
\]

Who Should I Cite: Learning Lit. Search Models From Cit. Behavior

- Retrieval model feature score: \( \text{score}(q,d) = \sum_i w_i \times f_i(q,d) \)
- 18 Features in 6 categories:
  - Similar Terms
  - Cited By Others
  - Recency
  - Cited Using Similar Terms
  - Similar Topics
  - Social Habits

Who Should I Cite: Learning Lit. Search Models From Cit. Behavior

- Apply Iterative Training Algorithm to build a retrieval model
- Types of Classifiers
  - Logistic classifier trained with quadratic prior
  - SVM-MAP classifier
- Results:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>(N = 100)</th>
<th>(N = 2000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>7.9</td>
<td>16.8</td>
</tr>
<tr>
<td>Logistic 50/50</td>
<td>19.9</td>
<td>24.9</td>
</tr>
<tr>
<td>SVM-MAP</td>
<td>25.3</td>
<td>27.9</td>
</tr>
</tbody>
</table>

Recommending Academic Papers Via Users’ Reading Purpose

- Main Components:
  - Abstract Splitting
  - Paper Similarity Model
    - TF*IDF Model
    - Topic Model
    - Concept-Based Topic Model
  - Candidate Selection
  - Paper Ranking

Table 1: NDCG result

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Problem Relevance</th>
<th>Solution Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF*IDF</td>
<td>0.8368</td>
<td>0.4910</td>
</tr>
<tr>
<td>LDA</td>
<td>0.8012</td>
<td>0.8336</td>
</tr>
<tr>
<td>LDA+Concept</td>
<td><strong>0.8403</strong></td>
<td><strong>0.8867</strong></td>
</tr>
</tbody>
</table>

Table 2: MAP result

<table>
<thead>
<tr>
<th>Model Name</th>
<th>Problem Relevance</th>
<th>Solution Relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF*IDF</td>
<td>0.8394</td>
<td>0.4483</td>
</tr>
<tr>
<td>LDA</td>
<td>0.7601</td>
<td>0.8313</td>
</tr>
<tr>
<td>LDA+Concept</td>
<td><strong>0.8427</strong></td>
<td><strong>0.8642</strong></td>
</tr>
</tbody>
</table>

Conclusion
Conclusion

• Content-Based Filtering Methods
  ○ Citation-Based Methods
    ■ Use existing citation graph to make recommendations
  ○ Context-Aware Methods
    ■ A subset of citation-based methods
    ■ Use words around a given citation to recommend papers to cite
  ○ Keyword-Based Methods
    ■ Analyze various components of paper (title and abstract) to make recommendations
Brief Description of Project

- Content-Based Filtering Method
  - Title, abstract, and body of user input paper
- Create a graph of papers in dataset
  - Edge exists if fused similarity metric between papers exceeds a specified threshold
- Compare input paper to those in graph
  - Rank via highest similarity based on fused metric
Questions?