ClusCite: Effective Citation Recommendation by Information Network-Based Clustering

Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, Jiawei Han
University of Illinois, at Urbana Champaign
27th August, 2014
Motivation

Personalized recommendation in social tagging systems using hierarchical clustering
A Shepitzen, J Gemmell, B Mobasher - Proceedings of the 2008 ... 2008 - dl.acm.org
... for profit or commercial advantage and that copies bear this notice and the full citation on the ... Other collaborative tagging applications focus on blogs, citations and wikis. ... If the personalized approach moves the resource further down the ranking in the recommendation set, the ...
Cited by 334 Related articles All 11 versions Cite Save

Context-aware citation recommendation
Q He, J Pei, D Kifer, P Mitra, L Giles - Proceedings of the 19th ... 2010 - dl.acm.org
... The bibliography candidates provided by a global recommendation should collectively satisfy the citation information needs of all out-link ... Definition 3.3 (Local Recommendation). ... out-link local context c+ with respect to d, a local recommendation is a ranked list of citations in a ...
Cited by 94 Related articles All 16 versions Cite Save

Citation recommendation without author supervision
Q He, D Kifer, J Pei, P Mitra, CL Giles - on Web search and data mining, 2011 - dl.acm.org
... SETUP In this section, we introduce notation and terminology, and describe the citation recommendation problem. ... it to try to recognize locations in the query manuscript d where citations should exist. ... goal is to cluster this bipartite graph to obtain clusters of citation contexts and ...
Cited by 28 Related articles All 10 versions Cite Save
Motivation

- Research papers need to cite relevant and important previous work
  - understand its background, context and innovation
- Already large, rapidly growing body of scientific literature
  - automatic recommendations of high quality citation
- Traditional literature search systems
  - infeasible to cast users’ rich information needs into queries consisting of just a few keywords
Citation Recommendation

Given a newly written manuscript (title, abstract and/or content) and its attributes (authors, target venues), suggests a small number of papers as high quality references.
Consider Distinct Citation Behavioral Patterns

**Basic idea:** Paper’s citations can be organized into different groups, each having its own behavioral pattern to identify references of interest

- A principled way to capture paper’s citation behaviors
- More accurate approach: paper-specific recommendation model

**Most existing studies:** assume *all* papers adopt *same criterion* and follow *same behavioral pattern* in citing other papers:

- **Context-based** [He et al., WWW’10; Huang et al., CIKM’12]
- **Topical similarity-based** [Nallapati et al., KDD’08; Tang et al., PAKDD’09]
- **Structural similarity-based** [Liben-Nowell et al., CIKM’03; Strohman et al., SIGIR’07]
- **Hybrid methods** [Bethard et al., CIKM’10; Yu et al., SDM’12]
Observation: Distinct Citation Behavioral Patterns

Each group follows distinct behavioral patterns and adopt different criterions in deciding relevance and authority of a candidate paper.
Observation: Distinct Citation Behavioral Patterns

Each group follow distinct behavioral patterns and adopt different criterions in deciding **relevance** and **authority** of a candidate paper.
Observation: Distinct Citation Behavioral Patterns

Each group follow distinct behavioral patterns and adopt different criterions in deciding relevance and authority of a candidate paper.
Observation: Distinct Citation Behavioral Patterns

Each group follow distinct behavioral patterns and adopt different criterions in deciding **relevance** and **authority** of a candidate paper.
Each group follows distinct behavioral patterns and adopt different criterions in deciding relevance and authority of a candidate paper.
Observation: Distinct Citation Behavioral Patterns

Each group follow distinct behavioral patterns and adopt different criterions in deciding **relevance** and **authority** of a candidate paper.
Heterogeneous Bibliographic Network

A unified graph representation for bibliographic dataset (papers and their attributes)

- Captures **paper-paper relevance** of different semantics
- Enables **authority propagation** between different types of objects
Given heterogeneous bibliographic network; terms, authors and target venues of a query manuscript \( q \), we aim to build a recommendation model specifically for \( q \), and recommend a small subset of papers as high quality references for \( q \).
We explore the principle that: citations tend to be softly clustered into different interest groups, based on the heterogeneous network structures.
We explore the principle that: citations tend to be *softly* clustered into different *interest groups*, based on the heterogeneous network structures.

For *different* interest groups, learn *distinct* models on finding relevant papers and judging authority of papers.

**Phrase I: Joint Learning (offline)**

**Derive group membership for query manuscript**

**Paper-specific recommendation model:** by integrating learned models of its related interest groups.

**Phrase II: Recommendation (online)**
Proposed Model: Overview

How likely a query manuscript $q$ will cite a candidate paper $p$ (suppose K interest groups):

$$s(q, p) = \sum_{k=1}^{K} \theta_q^{(k)} \cdot \left\{ r^{(k)}(q, p) + f_P^{(k)}(p) \right\}$$
Proposed Model: Overview

How likely a query manuscript $q$ will cite a candidate paper $p$ (suppose $K$ interest groups):

$$s(q, p) = \sum_{k=1}^{K} \theta^{(k)} \cdot \left\{ r^{(k)}(q, p) + f^{(k)}_P(p) \right\}$$

- query’s group membership
- relative citation score (how likely $q$ will cite $p$) within each group
Proposed Model: Overview

How likely a query manuscript $q$ will cite a candidate paper $p$ (suppose $K$ interest groups):

$$s(q, p) = \sum_{k=1}^{K} \theta^{(k)}_q \cdot \left[ r^{(k)}(q, p) + f^{(k)}_p(p) \right]$$

- **query’s group membership**
- **paper relative relevance** (query-candidate paper)
- **paper relative authority** (candidate paper)
Proposed Model: Overview

How likely a query manuscript $q$ will cite a candidate paper $p$ (suppose $K$ interest groups):

$$ s(q, p) = \sum_{k=1}^{K} \theta_q^{(k)} \cdot \left( r^{(k)}(q, p) + f_p^{(k)}(p) \right) $$

- query’s group membership
- paper relative relevance (query-candidate paper)
- paper relative authority (candidate paper)

It is desirable to suggest papers that have high relevance and authority scores across multiple related interest groups of the query manuscript.
Proposed Model I: Group Membership

- Learn each query’s group membership: scalability & generalizability
- Leverage the group memberships of related attribute objects to approximate query’s group membership

Different types of attribute objects ($X = \text{authors/venues/terms}$)

Query’s related (linked) objects of type-$X$

Attribute object’s group membership (to learn)
Proposed Model II: Paper Relevance

\[ s(q, p) = \sum_{k=1}^{K} \theta_q^{(k)} \cdot \left\{ r^{(k)}(q, p) + f_p^{(k)}(p) \right\} \]

\[ r^{(k)}(q, p) = \sum_{l=1}^{L} w_k^{(l)} \phi^{(l)}(q, p) \]

**Table 1: Meta paths with different semantics.**

<table>
<thead>
<tr>
<th>Meta path</th>
<th>Semantic meaning of the relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P - A - P )</td>
<td>( p_i ) and ( p_j ) share same author(s)</td>
</tr>
<tr>
<td>( P - T - P )</td>
<td>( p_i ) and ( p_j ) contain same term(s)</td>
</tr>
<tr>
<td>( P - V - P )</td>
<td>( p_i ) and ( p_j ) are in the same venue</td>
</tr>
<tr>
<td>( P - T - P \rightarrow P )</td>
<td>( p_i ) share term(s) with the paper(s) that cite ( p_j )</td>
</tr>
<tr>
<td>( P - A - P \leftarrow P )</td>
<td>( p_i ) share the same author(s) with the paper(s) cited by ( p_j )</td>
</tr>
</tbody>
</table>
Proposed Model II: Paper Relevance

Relevance features play different roles in different interest groups.

Weights on different meta path-based features.

Relevance features can be represented as:

\[
s(q, p) = \sum_{k=1}^{K} \theta^{(k)} \cdot \left\{ r^{(k)}(q, p) + f^{(k)}(p) \right\}
\]

Where:

- \( s(q, p) \) represents the relevance of paper \( p \) to query \( q \).
- \( \theta^{(k)} \) are weights.
- \( r^{(k)}(q, p) \) and \( f^{(k)}(p) \) are meta path-based features.

Weights on different meta paths:

<table>
<thead>
<tr>
<th>Meta path</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
<th>Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P - A - P )</td>
<td>0.0024</td>
<td>0.0113</td>
<td>0.0158</td>
<td>0.0765**</td>
</tr>
<tr>
<td>( P - A - P )</td>
<td>0.0054</td>
<td>0.0006</td>
<td>0.0192</td>
<td>0.1243</td>
</tr>
<tr>
<td>( P - A - P \rightarrow P )</td>
<td>0.6133**</td>
<td>0.2159*</td>
<td>0.2254</td>
<td>0.0213</td>
</tr>
<tr>
<td>( P - T - P )</td>
<td>0.1227</td>
<td>0.0947</td>
<td>0.1579</td>
<td>0.1095</td>
</tr>
<tr>
<td>( P - T - P \rightarrow P )</td>
<td>0.0442</td>
<td>0.5448**</td>
<td>0.3250*</td>
<td>0.0231</td>
</tr>
<tr>
<td>( P - T - P \leftarrow P )</td>
<td>0.1938*</td>
<td>0.0870</td>
<td>0.3578**</td>
<td>0.2409**</td>
</tr>
</tbody>
</table>

Table 1: Meta paths with different semantics.

<table>
<thead>
<tr>
<th>Meta path</th>
<th>Semantic meaning of the relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P - A - P )</td>
<td>( p_i ) and ( p_j ) share same author(s)</td>
</tr>
<tr>
<td>( P - T - P )</td>
<td>( p_i ) and ( p_j ) contain same term(s)</td>
</tr>
<tr>
<td>( P - V - P )</td>
<td>( p_i ) and ( p_j ) are in the same venue</td>
</tr>
<tr>
<td>( P - T - P \rightarrow P )</td>
<td>( p_i ) share term(s) with the paper(s) that cite ( p_j )</td>
</tr>
<tr>
<td>( P - A - P \leftarrow P )</td>
<td>( p_i ) share the same author(s) with the paper(s) cited by ( p_j )</td>
</tr>
</tbody>
</table>
Proposed Model III: Object Relative Authority

\[ s(q, p) = \sum_{k=1}^{K} \theta_{q}^{(k)} \cdot \left\{ r_{r}^{(k)}(q, p) + f_{p}^{(k)}(p) \right\} \]

Paper relative authority: A paper may have quite different visibility/authority among different groups, even if it is overall highly cited.
Proposed Model III: Object Relative Authority

\[ s(q, p) = \sum_{k=1}^{K} \theta_q^{(k)} \cdot \left\{ r^{(k)}(q, p) + f^{(k)}_{\mathcal{P}}(p) \right\} \]

Paper relative authority: A paper may have quite different visibility/authority among different groups, even it is overall highly cited.
Proposed Model III: Object Relative Authority

\[ s(q, p) = \sum_{k=1}^{K} \theta_q^{(k)} \cdot \left\{ r_q^{(k)}(q, p) + f_p^{(k)}(p) \right\} \]

**Paper relative authority:** A paper may have quite different visibility/authority among different groups, even it is overall highly cited

Table 6: Top-5 authority venues and authors from two example interest groups derived by ClusCite.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Venue (database and information system)</th>
<th>Author</th>
<th>Group I</th>
<th>Group II (computer vision and multimedia)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VLDB</td>
<td>Hector Garcia-Molina</td>
<td>0.0763</td>
<td>Richard Szeliski</td>
</tr>
<tr>
<td>2</td>
<td>SIGMOD</td>
<td>Christos Faloutsos</td>
<td>0.0653</td>
<td>ACM MM</td>
</tr>
<tr>
<td>3</td>
<td>TKDE</td>
<td>Elisa Bertino</td>
<td>0.0500</td>
<td>ICCV</td>
</tr>
<tr>
<td>4</td>
<td>CIKM</td>
<td>Dan Suciu</td>
<td>0.0488</td>
<td>CVPR</td>
</tr>
<tr>
<td>5</td>
<td>SIGKDD</td>
<td>H. V. Jagadish</td>
<td>0.0488</td>
<td>ECCV</td>
</tr>
</tbody>
</table>

Relative authority propagation over the network.

**Network:** WSDM, KDD
A joint optimization problem:

\[
\min_{P, W, F_P, F_A, F_Y} \frac{1}{2} \mathcal{L} + \mathcal{R} + \frac{c_p}{2} \|P\|_F^2 + \frac{c_w}{2} \|W\|_F^2 \\
\text{s.t. } P \geq 0; \quad W \geq 0.
\]

Weighted model prediction error

Graph regularization for encoding authority propagation
Model Learning

A joint optimization problem:

$$\min_{P, W, F_P, F_A, F_Y} \frac{1}{2} \mathcal{L} + \mathcal{R} + \frac{c_P}{2} \|P\|_F^2 + \frac{c_w}{2} \|W\|_F^2$$

s.t. $P \geq 0; \quad W \geq 0.$

\begin{align*}
\mathcal{L} &= \sum_{i,j=1}^{n} M_{ij} \left( Y_{ij} - \sum_{k=1}^{K} \sum_{l=1}^{L} \theta_{p_i}^{(k)} w_{k}^{(l)} S_{jl}^{(i)} - \sum_{k=1}^{K} \theta_{p_i}^{(k)} F_{P,kj} \right)^2 \\
&= \sum_{i=1}^{n} \left\| M_i \odot \left( Y_i - R_i P (W S^{(i)T} + F_P) \right) \right\|_2^2.
\end{align*}

\begin{align*}
\mathcal{R} &= \frac{\lambda_A}{2} \sum_{i=1}^{n} \sum_{j=1}^{A} R_{ij}^{(A)} \left\| \frac{F_{P,i}}{D_{ii}^{(P,A)}} - \frac{F_{A,i}}{D_{jj}^{(A,P)}} \right\|_2^2 \\
&+ \frac{\lambda_B}{2} \sum_{i=1}^{n} \sum_{j=1}^{|\mathcal{V}|} R_{ij}^{(V)} \left\| \frac{F_{P,i}}{D_{ii}^{(P,V)}} - \frac{F_{V,i}}{D_{jj}^{(V,P)}} \right\|_2^2.
\end{align*}

Algorithm: alternating minimization (w.r.t. each variable)
Experimental Results

- **Datasets**
  - DBLP: 137k papers; ~2.3M relationships; Avg # citations/paper: 5.16
  - PubMed: 100k papers; ~3.6M relationships; Avg # citations/paper: 17.55

- **Case study on citation behavioral patterns**

Each paper is assigned to the group with the highest group membership score.
Experimental Results

- **Performance Comparisons**
  - 17.68% improvement in Recall@50; 9.57% in MRR, on DBLP

<table>
<thead>
<tr>
<th>Method</th>
<th>P@10</th>
<th>P@20</th>
<th>R@20</th>
<th>R@50</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>BM25</td>
<td>0.126</td>
<td>0.090</td>
<td>0.143</td>
<td>0.214</td>
<td>0.410</td>
</tr>
<tr>
<td>PopRank</td>
<td>0.011</td>
<td>0.009</td>
<td>0.015</td>
<td>0.030</td>
<td>0.045</td>
</tr>
<tr>
<td>TopicSim</td>
<td>0.032</td>
<td>0.027</td>
<td>0.043</td>
<td>0.082</td>
<td>0.116</td>
</tr>
<tr>
<td>Link-PLSA-LDA</td>
<td>0.102</td>
<td>0.089</td>
<td>0.129</td>
<td>0.182</td>
<td>0.374</td>
</tr>
<tr>
<td>L2-LR</td>
<td>0.227</td>
<td>0.167</td>
<td>0.247</td>
<td>0.354</td>
<td>0.486</td>
</tr>
<tr>
<td>RankSVM</td>
<td>0.237</td>
<td>0.179</td>
<td>0.273</td>
<td>0.362</td>
<td>0.499</td>
</tr>
<tr>
<td>MixFea</td>
<td>0.226</td>
<td>0.168</td>
<td>0.247</td>
<td>0.363</td>
<td>0.500</td>
</tr>
<tr>
<td>ClusCite-Rel</td>
<td>0.240</td>
<td>0.187</td>
<td>0.286</td>
<td>0.401</td>
<td>0.516</td>
</tr>
<tr>
<td>ClusCite</td>
<td><strong>0.243</strong></td>
<td><strong>0.196</strong></td>
<td><strong>0.299</strong></td>
<td><strong>0.428</strong></td>
<td><strong>0.548</strong></td>
</tr>
</tbody>
</table>
Experimental Results

- Performance Analysis
  - a) Number of attributes each query manuscript has

![Graph](a) # attribute objects
Experimental Results

- **Performance Analysis**
  a) Number of attributes each query manuscript has
  b) Number of interest groups

(a) # attribute objects
(b) PubMed
Conclusion

- Study citation recommendation in the context of heterogeneous bibliographic networks
- ClusCite framework: organize paper citations into interest groups, and design cluster-based models, yielding paper-specific recommendation
- Experimental results demonstrate significant improvement over state-of-the-art methods
Conclusion

- Study citation recommendation in the context of heterogeneous bibliographic networks

- ClusCite framework: organize paper citations into interest groups, and design cluster-based models, yielding paper-specific recommendation

- Experimental results demonstrate significant improvement over state-of-the-art methods

- Future work: the principle of clustering-based recommendation can be applied to other recommendation problems, e.g., movie recommendation; news recommendation.
Thank You!

Questions?
Figure 3: Correlation between paper relative authority and # ground truth citations, during different iterations.
Learning Algorithm

- **Learning algorithm**: alternative minimization w.r.t each variable
  - **Convergence**:
    - Guaranteed by an analysis similar to block coordinate descent
    - Empirically converges in 50 iterations
  
- **Mutually enhancing**:
  - Authority/relevance features can be better learned with high-quality groups derived
  - Good features help deriving high-quality groups.