Exploring The Leading Authors and Journals in Major Topics by Citation Sentences and Topic Modeling

Ha Jin Kim, Juyoung An, Yoo Kyung Jeong, and Min Song
Department of Library and Information Science, Yonsei University
Exploring intellectual structures

- **Author**: Collaboration, Author co-citation analysis, h-index, …
- **Journal**: Journal Impact Factor, SJR…
- **Content (Document, Research discipline)**: Document citation analysis, Co-word analysis, Topic models…
- **etc…**
Exploring intellectual structures
The Leading Author, Journal and Topic

- **Citation sentence:** Containing brief content of cited work and opinion that the author of citing work on the cited work

- **Topic Model:** Adopting Author Conference Topic (ACT) model (Tang, Jin & Zhang, 2008)

- **Oncology:** The recent surge in number of publications in this field. “Stem cells,” one of the subfields of oncology, has been at the forefront of medicine
Citation Sentence

Brief content of cited work and opinion that the author of citing work on the cited work

- Embedding useful contents signifying the influence of cited authors on shared ideas
- Being considered as an invisible intellectual place for idea exchanging
- Playing a role of supporting and expressing their own arguments by using other works

Exploring the implicit topics resided in citation sentences
Original ACT Model (Tang, Jin & Zhang, 2008)

Purpose of Academic search

AMiner: https://aminer.org/
Modified AJT Model

1) Citation Data Extraction

- Full-text Data collection
- PMC

Authors

2) Parameter Optimization

• Which topic is most salient?
• Who is the active authors sharing other authors’ ideas?
• Which journal leads such endeavor?
Method
Citation Sentence Extraction

1) Journal Title
2) Citation Sentences
3) Citing Authors

The 77-SNP PRS was associated with a larger effect than previously reported for a 10-SNP-PRS (\(\text{xref rid="CIT0020" ref-type="bibr"} \) 20 \(\text{xref}\)).
Data collection
Oncology field

- **PubMed Central**: 6,360 full-text articles
- **15 journals of Oncology**: by Thomson Reuter’s JCR & journal’s impact factor

*Cancer Cell, Journal of the National Cancer Institute, Leukemia, Oncogene, Annals of Oncology, Neuro-Oncology, Stem Cells, Oncotarget, OncoImmunology, Molecular Oncology, Breast Cancer Research Journal of Thoracic Oncology, Pigment Cell & Melanoma Resaerch, Clinical Epigenetics, Molecular Cancer*
1) Citation Data Extraction

2) Parameter Optimization

Topic 2
- 1st author
- 2nd author
- 1st journal
- 2nd journal

Research Flow

Full-text Data collection

AJT Model

Authors

Citation sentences

Journal Title

Context Identification of Sentences in Related Work Sections using a Conditional Random Field: 

M.A. Angioli, S. Chevalier, N. Sanger

1) Citation Data Extraction

Research Flow

Full-text Data collection

AJT Model

Authors

Citation sentences

Journal Title

Context Identification of Sentences in Related Work Sections using a Conditional Random Field: 

M.A. Angioli, S. Chevalier, N. Sanger
# Results (8 Topics)

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td>Cancer epigenetics</td>
<td>Leukemia</td>
<td>Targeted therapy</td>
</tr>
<tr>
<td>breast expression mammary risk women</td>
<td>methylation DNA expression gene histone</td>
<td>expression mutations AML treatment leukemia</td>
<td>mutations clinical treatment survival resistance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author Group 1</th>
<th>Author Group 2</th>
<th>Author Group 3</th>
<th>Author Group 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Johnston Stephen RD Colditz Graham A Sternlicht Mark D</td>
<td>Gray Steven G Mahlknecht Ulrich Tollefsbol Trygve O</td>
<td>Tefferi A Anderson K C Ratajczak Janina</td>
<td>Muller Patricia AJ Vousden Karen H Zaravinos Apostolos</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Journal Group 1</th>
<th>Journal Group 2</th>
<th>Journal Group 3</th>
<th>Journal Group 4</th>
</tr>
</thead>
</table>
## Results (cont’d)

<table>
<thead>
<tr>
<th>Topic 5</th>
<th>Topic 6</th>
<th>Topic 7</th>
<th>Topic 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular cancer</td>
<td>Oncogene pathway</td>
<td>Cancer Immunology</td>
<td>Stem Cell</td>
</tr>
<tr>
<td>expression</td>
<td>cell</td>
<td>cell</td>
<td>stem</td>
</tr>
<tr>
<td>p53 mutant</td>
<td>activity</td>
<td>immune</td>
<td>expression</td>
</tr>
<tr>
<td>gene survival</td>
<td>activation</td>
<td>expression</td>
<td>differentiation</td>
</tr>
<tr>
<td></td>
<td>protein</td>
<td>clinical</td>
<td>MSCs</td>
</tr>
<tr>
<td></td>
<td>apoptosis</td>
<td>responses</td>
<td>growth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author Group 5</th>
<th>Author Group 6</th>
<th>Author Group 7</th>
<th>Author Group 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarke Paul A</td>
<td>Melino Gerry</td>
<td>Zitvogel Laurence</td>
<td>Romagnani Paola</td>
</tr>
<tr>
<td>Workman Paul</td>
<td>Martelli Alberto M</td>
<td>Galluzi Lorenzo</td>
<td>Salem Husein K</td>
</tr>
<tr>
<td>Hoelder Swen</td>
<td>McCubrey James A</td>
<td>Kroemer Guido</td>
<td>Thiemermann Chris</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Journal Group 5</th>
<th>Journal Group 6</th>
<th>Journal Group 7</th>
<th>Journal Group 8</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Cancer Cell</em></td>
<td>Oncotarget</td>
<td><em>Oncoimmunology</em></td>
<td><em>Stem Cells</em></td>
</tr>
<tr>
<td>Neuro-Oncology</td>
<td>Annals of Oncology</td>
<td>Breast Cancer Research</td>
<td></td>
</tr>
<tr>
<td>Oncotarget</td>
<td>Cancer Cell</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Visualization of Topic Words

- **Topic 1**: Breast Cancer
- **Topic 2**: Clinical Epigenetics
- **Topic 3**: Leukemia
- **Topic 4**: Oncoimmunology
- **Topic 5**: Molecular Cancer
- **Topic 6**: Oncotarget
- **Topic 7**: Oncoimmunology
- **Topic 8**: Stem Cells (Dayton, Ohio)

General words on Oncology field

- # nodes: 100
- # edges: 1,436
- Nodes’ size: Degree centrality
Conclusion

- Exploration of Oncology field by adopting AJT model and using citation sentences
  - **AJT model**: to detect leading authors and journals in sub-disciplines represented by discovered topics in a certain field
  - **Citation sentences**: Discovering latent meaning associated citation sentences and the major players leading the field
Future works

• Comparing the proposed approach with the general topic modeling technique
• Investigating whether there is a different impact of using citation sentences and general meta-data (abstract and title)
• Considering the window size of citation sentences enriching citation context
Thank you!