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

TSTM
Text & Social Media Mining

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

Tang, J., Jin, R., & Zhang, J. (2008, December). A topic modeling approach and its integration into the random walk framework for academic search. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* (pp. 1055-1060). IEEE.

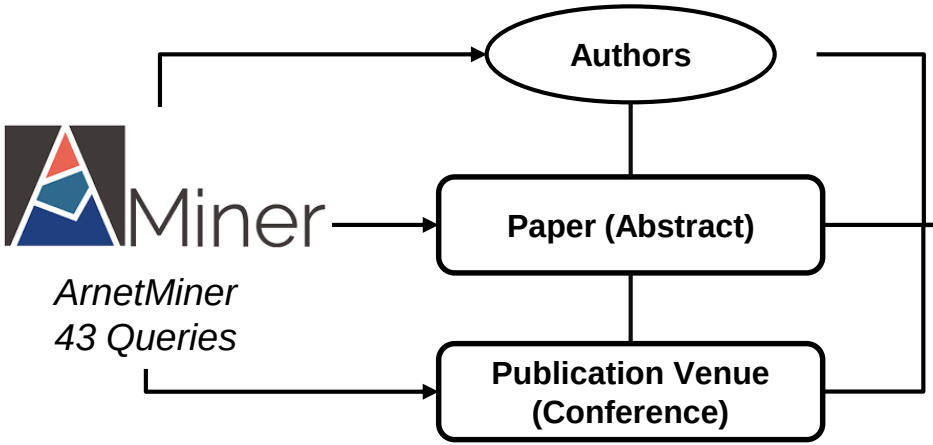
Citation Sentence

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)



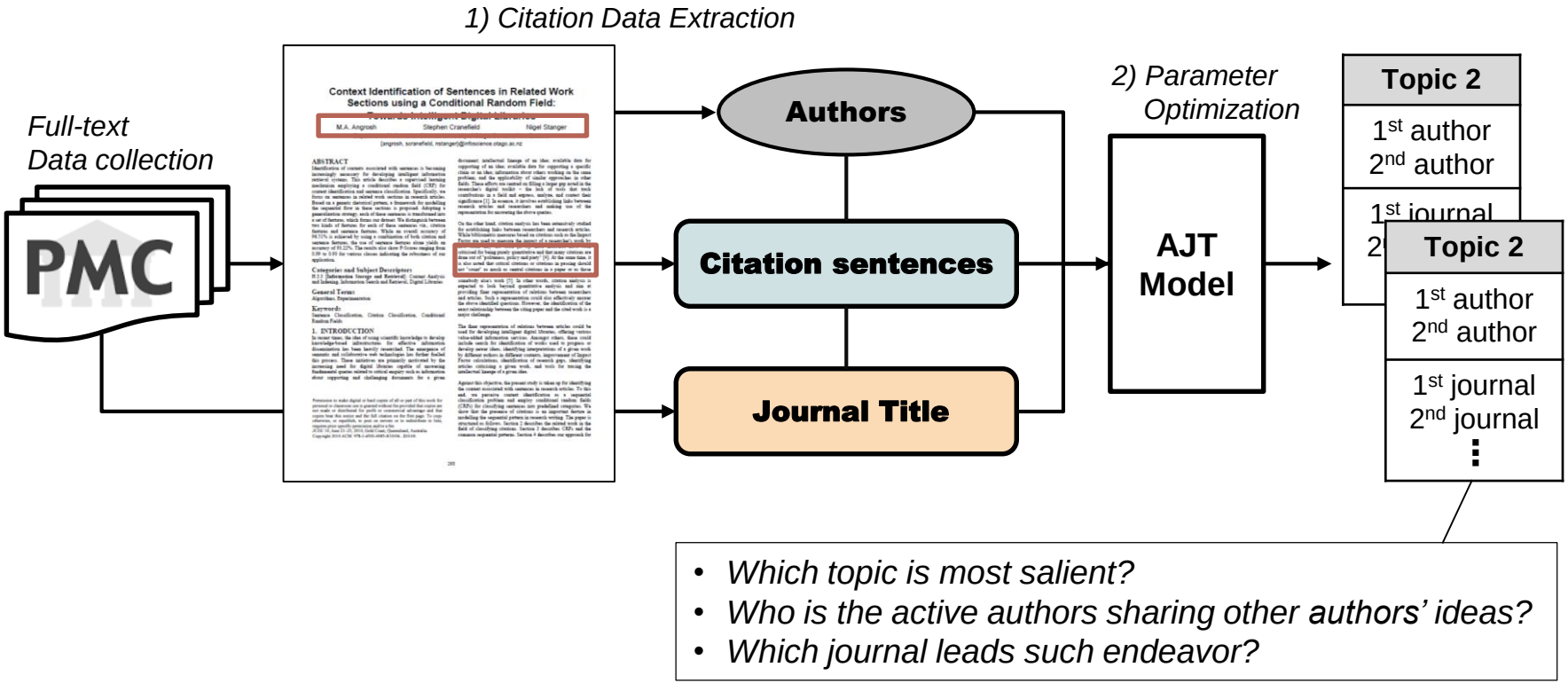
ACT & Random Walk

Method	Object	P@5	P@10	P@20	R-pre	MAP
BM25	Paper	42.9	45.7	41.4	12.0	47.2
	Author	77.1	47.1	26.4	67.5	85.5
	Conference	51.4	38.6	22.9	48.8	66.0
	Average	57.1	43.8	30.2	42.8	66.2
LM	Paper	40.0	38.6	37.1	10.0	46.4
	Author	65.7	44.3	25.0	58.8	73.4
	Conference	51.4	32.9	21.4	47.6	63.1
	Average	52.4	38.6	27.9	38.8	61.0
LM+RW1	Paper	62.9	55.7	44.3	12.9	65.3
	Author	71.4	48.6	25.7	64.6	83.8
	Conference	60.0	35.7	22.1	53.6	64.6
	Average	64.8	46.7	30.7	43.7	71.2
pLSI	Paper	32.5	33.8	30	9.7	40.4
	Author	65.0	40.0	22.5	60.4	75.5
	Conference	47.5	36.3	21.3	45.1	54.1
	Average	48.3	36.7	24.6	38.4	56.7
LDA	Paper	31.4	48.6	42.9	13.5	45.8
AT	Paper	42.9	48.6	42.9	13.1	49.3
	Author	82.9	45.7	25.7	73.5	78.1
	Average	62.9	47.1	34.3	43.3	63.7
ACT1	Paper	42.9	45.7	43.6	16.6	51.0
	Author	91.4	50.0	26.4	80.0	89.6
	Conference	62.9	41.4	23.6	60.7	72.3
	Average	65.7	45.7	31.2	52.4	71.0
ACT1+RW1	Paper	68.6	61.4	50.7	17.1	66.6
	Author	80.0	51.4	27.1	77.6	87.4
	Conference	62.9	42.9	23.6	59.5	72.0
	Average	**70.5**	**51.9**	**33.8**	**51.4**	**75.4**

Purpose of Academic search

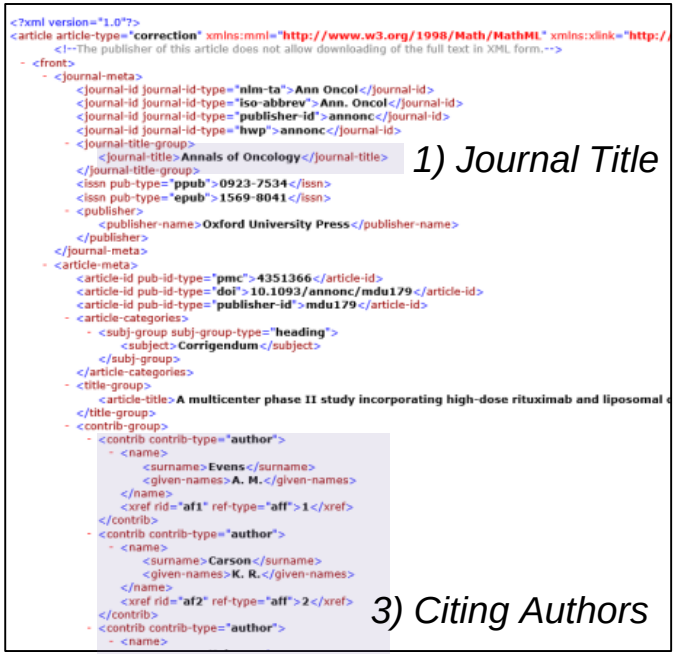
AMiner: https://aminer.org/

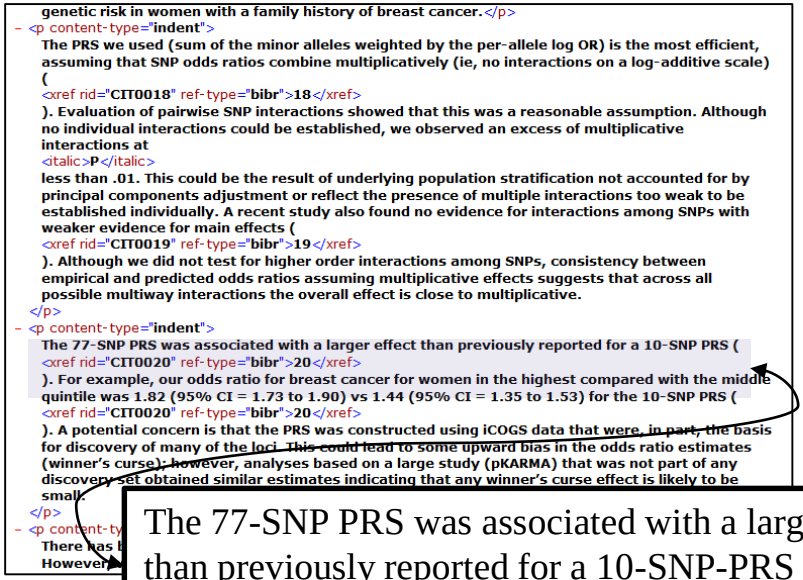
Modified AJT Model



Method

Citation Sentence Extraction





The 77-SNP PRS was associated with a larger effect than previously reported for a 10-SNP-PRS **(<xref rid="CIT0020" ref-type="bibr"> 20 </xref>).**

2) Citation Sentences

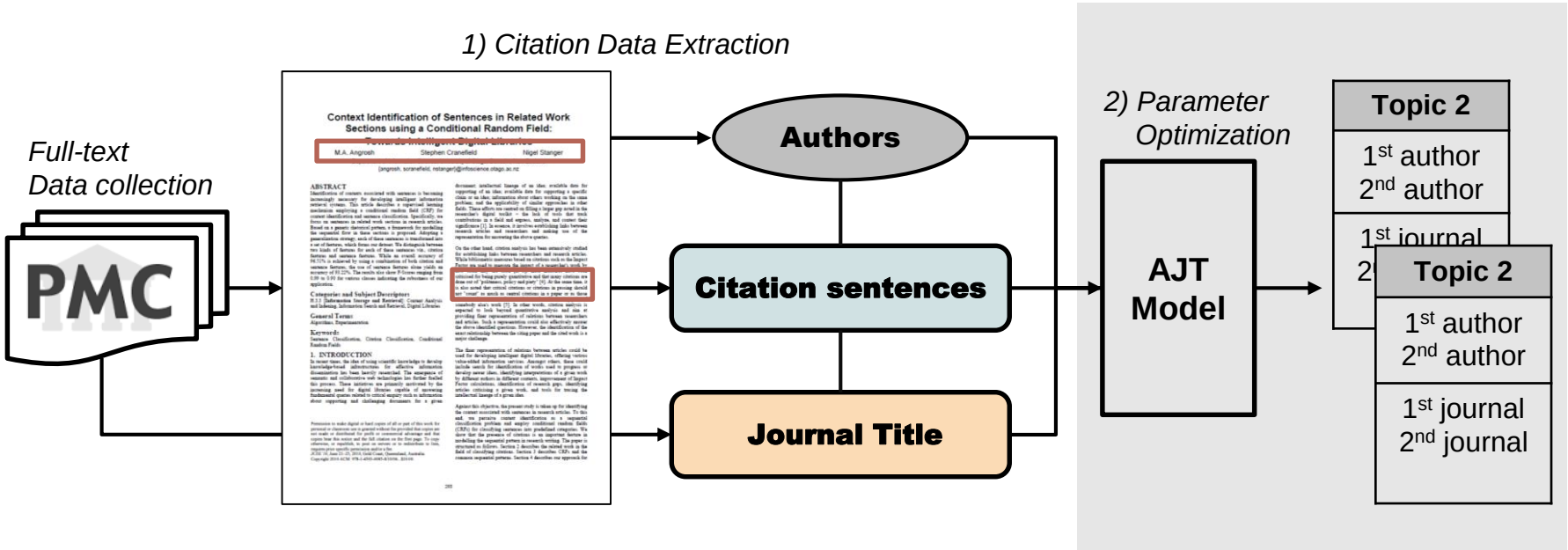
Data collection

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, OncoInnunology, Molecular Oncology, Breast Cancer Research Journal of Thoracic Oncology, Pigment Cell & Melanoma Resaerch, Clinical Epigenetics, Molecular Cancer

Research Flow



Results (8 Topics)

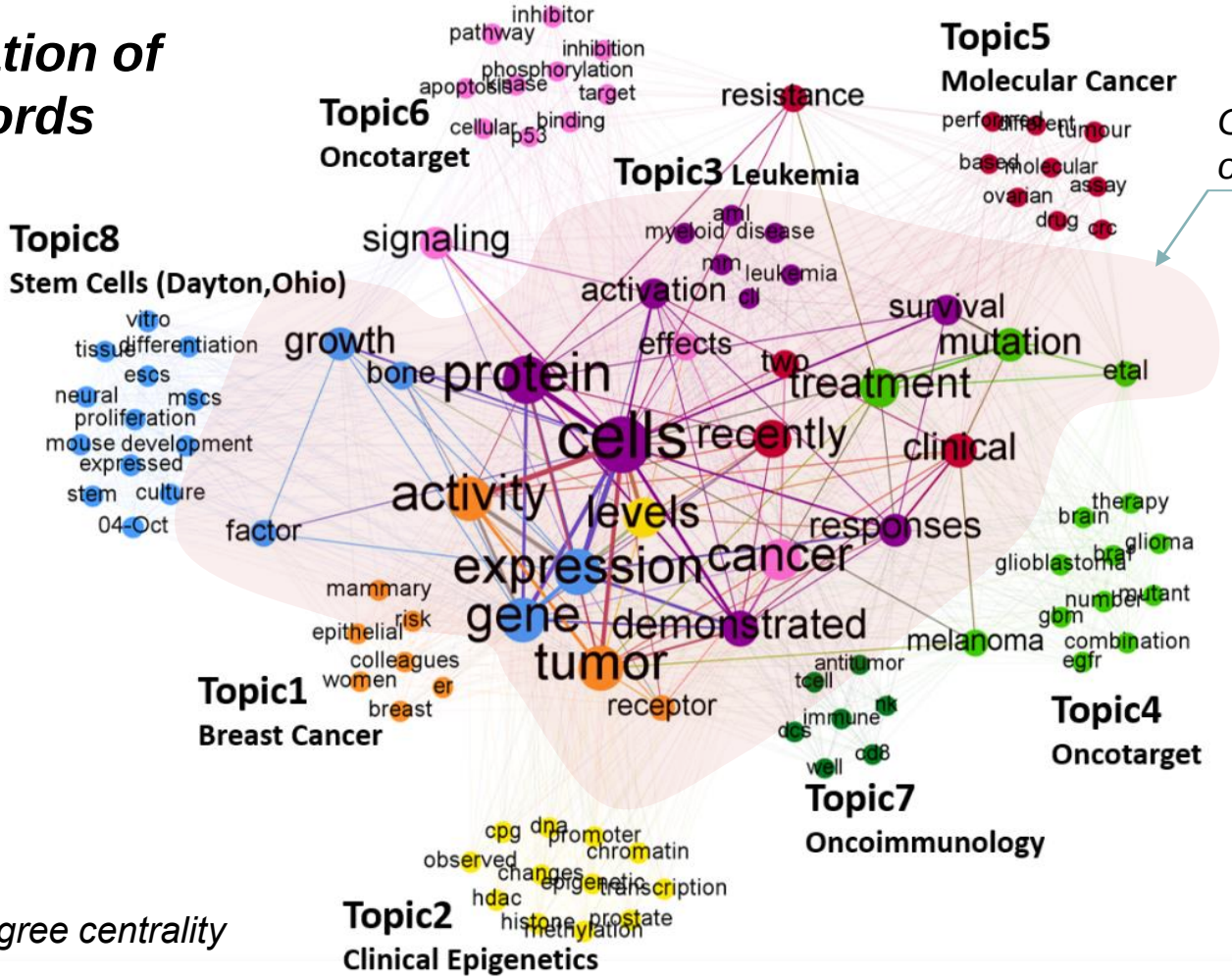
Labeled by 3 Experts

Topic 1 Breast cancer	**Topic 2** Cancer epigenetics	**Topic 3** Leukemia	**Topic 4** Targeted therapy
breast expression mammary risk women	methylation DNA expression gene histone	expression mutations AML treatment leukemia	mutations clinical treatment survival resistance
Author Group 1	**Author Group 2**	**Author Group 3**	**Author Group 4**
Johnston Stephen RD Colditz Graham A Sternlicht Mark D	Gray Steven G Mahlknecht Ulrich Tollefsbol Trygve O	Tefferi A Anderson K C Ratajczak Janina	Muller Patricia AJ Vousden Karen H Zaravinos Apostolos
Journal Group 1	**Journal Group 2**	**Journal Group 3**	**Journal Group 4**
Breast Cancer Research Annals of Oncology	*Clinical Epigenetics* Oncoimmunology	*Leukemia* Pigment Cell & Melanoma Research	*Oncotarget* Journal of Thoracic Oncology

Results (cont'd)

Topic 5 Molecular cancer	**Topic 6** Oncogene pathway	**Topic 7** Cancer Immunology	**Topic 8** Stem Cell
expression	cell	cell	stem
p53	activity	immune	expression
mutant	activation	expression	differentiation
gene	protein	clinical	MSCs
survival	apoptosis	responses	growth
Author Group 5	**Author Group 6**	**Author Group 7**	**Author Group 8**
Clarke Paul A	Melino Gerry	Zitvogel Laurence	Romagnani Paola
Workman Paul	Martelli Alberto M	Galluzzi Lorenzo	Salem Husein K
Hoelder Swen	McCubrey James A	Kroemer Guido	Thiemermann Chris
Journal Group 5	**Journal Group 6**	**Journal Group 7**	**Journal Group 8**
Cancer Cell	*Oncotarget*	*Oncoimmunology*	*Stem Cells*
Neuro-Oncology	Annals of Oncology	Annals of Oncology	Annals of Oncology
Oncotarget	Cancer Cell	Breast Cancer Research	Cancer Cell

Visualization of Topic Words



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!