

# Text Mining for Research Portfolio Analysis: Automatically Cataloging Related Journals in Biomedicine

Yuqing Mao<sup>1</sup> and Zhiyong Lu<sup>1,\*</sup>

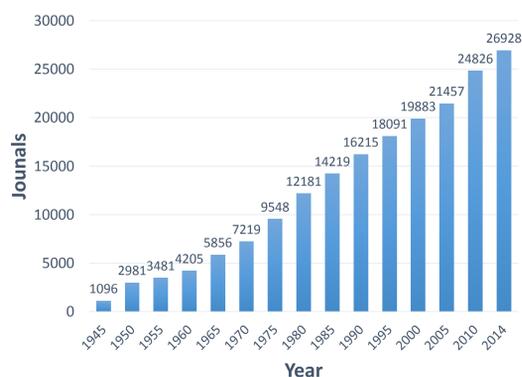
<sup>1</sup>National Center for Biotechnology Information (NCBI), National Library of Medicine, Bethesda, MD 20894

Contact: zhiyong.lu@nih.gov

## Introduction

Clustering journals of similar topics is important for research portfolio analysis (e.g. grouping and sorting journals in the same category for assessing the significance of research).

Traditionally, journal clustering is based on human cataloging. However, this manual approach is unable to *a*) keep up with the rapid growth of new journals, *b*) capture the changes in journal scopes over time, and *c*) measure the relatedness between journals. Here, we present data-driven approaches for automatically identifying related journals in a timely fashion.



Number of journals in PubMed From 1945-2014

## Methods

### Citation-based:

- If articles in two journals often cite each another, then the two journals are likely to be related.
- Calculating citation scores for journals that give or receive the highest number of citations (maximal of “to” or “from” the journal) [1].

### Usage-based:

- If articles in two journals are often seen by the same set of users, then the two journals are likely to be related.
- PubMed query log data (searches and clicks) can be approximated as measures of article usage [2].

### Content-based:

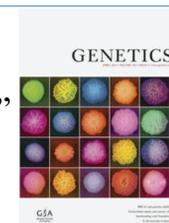
- If articles in two journals often publish similar topics, then the two journals are likely to be related.
- Similarity between articles can be measured based on their content using term weights [3].

### Data:

- PubMed articles published from August 1, 2011 to July 31, 2012
- 917,844 articles belonging to 4,841 unique journals
- One month’s PubMed query logs: Millions of user sessions and citation clicks.

## Results

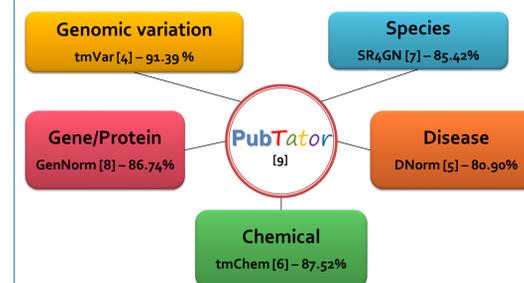
Input journal ex: “Genetics”  
Top 20 related journals based on three methods



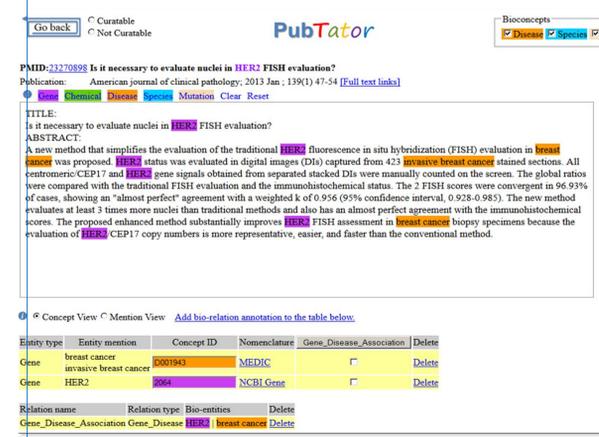
Citation-based	Usage-based	Content-based
Proceedings of the National Academy of Sciences of the USA	Proceedings of the National Academy of Sciences of the USA	PLoS Genetics
Cell	Journal of Biological Chemistry	Proceedings of the National Academy of Sciences of the USA
Nature	Nature	Molecular and Cellular Biology
Molecular and Cellular Biology	Science	PLoS One
Science	Molecular and Cellular Biology	Genome Research
Theoretical and Applied Genetics	Cell	Nature
Evolution	Development	Eukaryotic Cell
EMBO journal	Genes & Development	Evolution
Genes & Development	Nucleic Acid Research	Molecular Biology of the Cell
Nucleic Acid Research	EMBO journal	Molecular Ecology
Journal of Biological Chemistry	Current Biology	Journal of Biological Chemistry
Molecular Biology and Evolution	Molecular Biology of the Cell	Science
Journal of Bacteriology	Developmental Biology	Developmental Biology
Molecular & General Genetics	Molecular Biology and Evolution	Theoretical Population Biology
Heredity	Nature Genetics	Current Biology
Development	Journal of Cell Biology	Genes & Development
Genetical Research	Journal of Bacteriology	Development
Genome	Critical care medicine	Molecular Biology and Evolution
J. of Molecular Biology	PLoS Genetics	Heredity
Journal of Cell Biology	American Journal of Human Genetics	American Journal of Human Genetics

## Other Text Mining Tools

We have developed state-of-the-art software tools for recognizing key biomedical concepts in free text: gene/proteins, diseases, chemicals, species, cell lines, and mutations.



Further, we have applied our tools to all PubMed articles and integrated text-mined results into PubTator: a web-based application for assisting literature curation.



Entity type	Entity mention	Concept ID	Nomenclature	Gene_Disease_Association	Delete
Gene	breast cancer	0501943	MEDIC	<input type="checkbox"/>	Delete
Gene	HER2	2064	NCBI Gene	<input type="checkbox"/>	Delete

## References

1. Pudovkin AI, Garfield E. Algorithmic procedure for finding semantically related journals. Journal of the American Society for Information Science and Technology. 2002;53(13):1113-9.
2. Lu Z, Xie N, Wilbur WJ. Identifying related journals through log analysis. Bioinformatics. 2009;25(22):3038-9
3. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. BMC bioinformatics. 2007;8(1):423.
4. Wei CH, Harris BR, Kao HY, Lu Z, tmVar: A text mining approach for extracting sequence variants in biomedical literature, 2013, Bioinformatics, 29(11) 1433–1439
5. Leaman R, Doğan RI and Lu Z, DNORM: Disease Name Normalization with Pairwise Learning to Rank, 2013, Bioinformatics, 29 (22): 2909-2917
6. Leaman R, Wei CH, Lu Z, NCBI at the BioCreative IV CHEMDNER Task: Recognizing chemical names in PubMed articles with tmChem, BioCreative IV, 2013, vol 2, 34-41
7. Wei CH, Kao HY, Lu Z, SR4GN: a species recognition software tool for gene normalization, 2012, PLoS ONE, 7(6):e38460
8. Wei CH, Kao HY (2011) Cross-species gene normalization by species inference. BMC Bioinformatics, 12(Suppl 8):S5
9. Wei CH et. al., PubTator: a Web-based text mining tool for assisting Biocuration, Nucleic acids research, 2013, 41 (W1): W518-W522

## Acknowledgments

Funding: This research is supported by NIH Intramural Research Program, National Library of Medicine.

More information at <http://goo.gl/sdRf3Y>

