

openMINDED

TDM in recommender systems for research and in tracking research impact

Petr Knoth
Knowledge Media institute, The Open
University



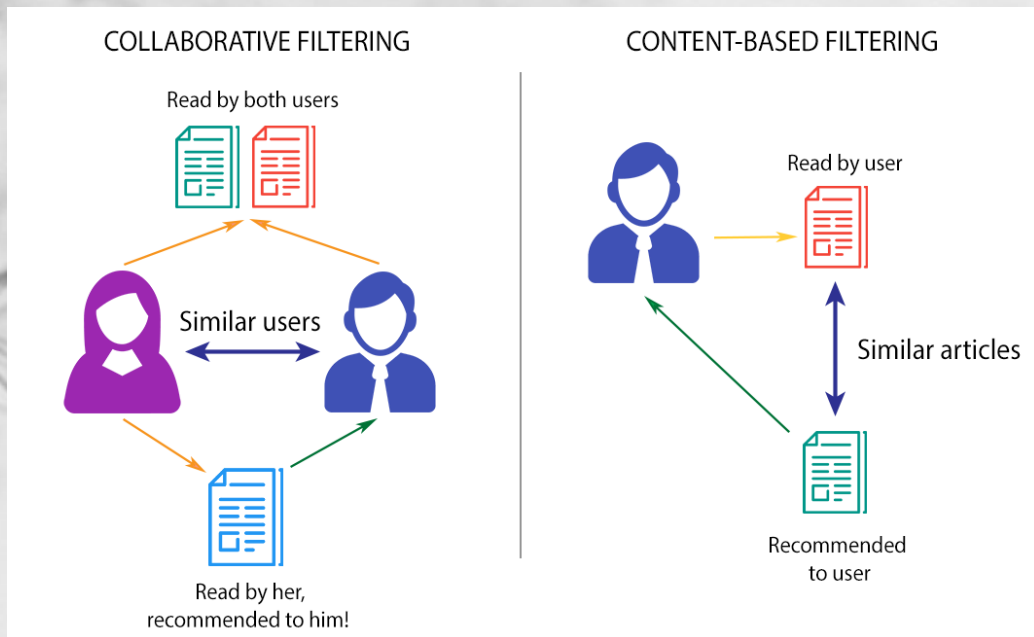
United Kingdom



TDM in recommender systems for research

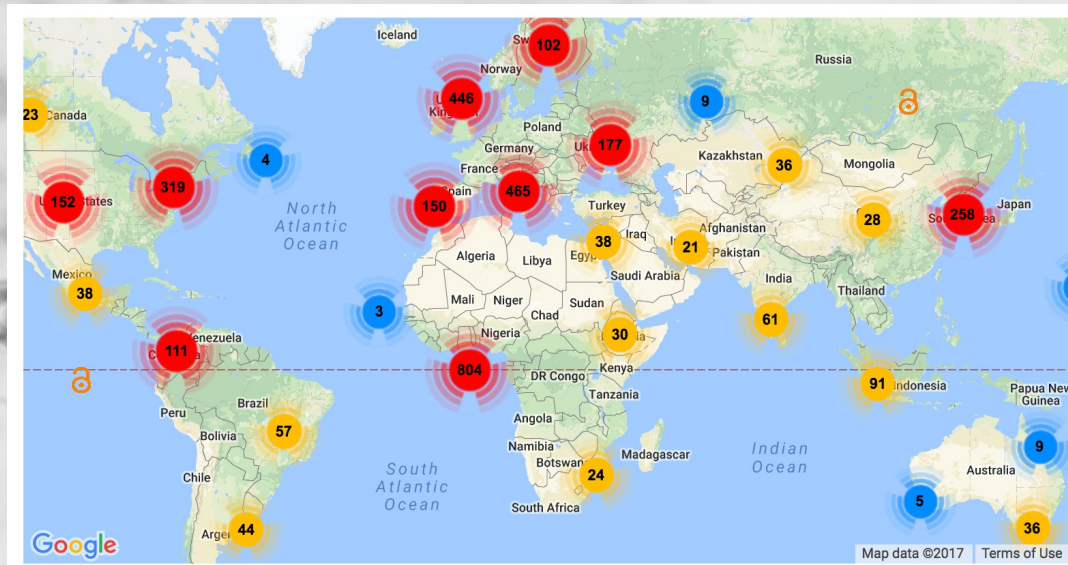
Why TDM in recommender systems for research?

- Collaborative filtering vs content-based filtering
- In the scholarly databases, we have many documents but relatively few users => content-based filtering
- Recommending entities



The CORE recommender system

- CORE provides a content-based recommendation system for articles from across the global network of repositories.
- Dataset:
 - 8.3 million full texts
 - 79 million metadata records
 - 3,658 data providers



Recommendation as a service

The screenshot displays the Apollo digital repository interface. At the top, the University of Cambridge logo is visible. Below it, a navigation bar includes links to 'Apollo Home', 'School of the Physical Sciences', 'Department of Chemistry', 'Unilever Centre for Molecular Informatics', 'Panton Discussions', and 'View Item'. The Apollo logo and a search bar are also present. The main content area shows the record for 'Open Content Mining'. On the left, a sidebar lists navigation options: 'All of Apollo' (Communities & Collections, Authors, Titles, Keywords, Type), 'This Collection' (Authors, Titles, Keywords, Type), 'Statistics' (View Usage Statistics), and 'Browse'. The main record details include: 'View / Open Files' (article text PDF, 9Mb), 'Authors' (Murray-Rust, Peter), 'Publication Date' (2012-09-24), 'ISBN' (to be assigned), 'Language' (English), 'Citation' (Murray-Rust, P. (2012). Open Content Mining.), 'Description' (Conference for the Fellows of OpenForum Academy - 24th September 2012 Brussels), 'Abstract' (We present evidence that content-mining of scholarly articles is now technically feasible and highly valuable both. However researchers and information technologists are blocked by legal and contractual barriers from using it and developing the methodologies. We review the problems and propose changes in legal policy which we have already submitted to the UK's Hargreaves report on intellectual property reform. We put forward the fundamental rights of scholars and embed them in a manifesto: "The right to read is the right to mine", "Users and providers should encourage machine processing, and "Facts don't belong to anyone".), 'Keywords' (Open Content Mining, Index Terms—Open Knowledge, Content mining, Hargreaves process, Text mining, publishers, legal barriers), and 'Identifiers' (This record's URL: http://www.dspace.cam.ac.uk/handle/1810/243749).

- Recommender plugin for repositories
- Recommendations from the CORE API

Recommendation as a service

- Recommender plugin for repositories
- Recommendations from the CORE API

Type
Conference Object


Metadata
[Show full item record](#)


Rights
Attribution 2.0 UK: England & Wales
Licence URL:
<http://creativecommons.org/licenses/by/2.0/uk/>


Recommended or similar items


Suggested articles


Suggested articles in Apollo


**Effectively and Efficiently Mining Frequent Patterns from Dense Graph Streams on Disk**
Provided by: Elsevier - Publisher Connector | **Publisher:** The Authors. Published by Elsevier B.V. | **Year:** 2014
By Braun Peter, Cameron Juan J., Cuzzocrea Alfredo, Jiang Fan, Leung Carson K.

**The right to read is the right to mine: Text and data mining copyright exceptions introduced in the UK.**
Provided by: LSE Research Online | **Publisher:** London School of Economics and Political Science | **Year:** 2014
By Mounce Ross

**Global boom, local impacts: Mining revenues and subnational outcomes in Peru 2007-2011**
Provided by: EconStor | **Publisher:** Washington, DC: Inter-American Development Bank (IDB) | **Year:** 2014
By Zambrano Omar, Robles Marcos, Laos Denise

**Environmental security, mining and good governance : mining regulation in the Kyrgyz region. A review**
Provided by: UEF Electronic Publications | **Publisher:** University of Eastern Finland
By Honkonen T

**A Case Study of Data Analysis Process and Tools for a Consulting Company**
Provided by: Aaltodoc Publication Archive | **Year:** 2012
By Gong Peng

Powered by  CORE

How does the CORE recommender system work?

Article-article recommender system. Processes:

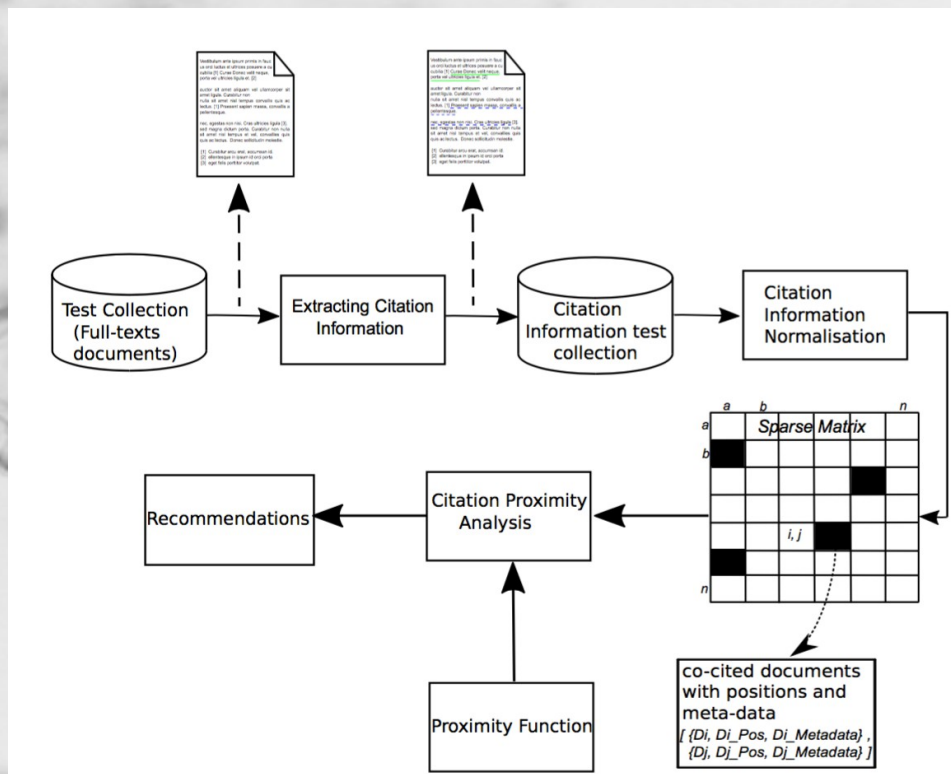
1. Preprocessing prior to recsys: feature extraction/enrichment with e.g. document type, citation and citation proximity data, identifiers, etc.
2. Similarity measure/ranking function
3. Post-filtering using record quality
4. Feedback (crowdsourcing a black list)

Combining features

- Evaluating different ranking functions (P,R,MAP, etc.):
 - Weights for boosting
 - Scaling function (e.g. exponential decay for recency)
- Offline ground truths:
 - MAG citation assumption
 - MAG co-citation assumption
- Learning to rank (haven't done yet)
- Online A/B testing (haven't done yet)

Citation proximity analysis

- CPA extends the co-citation assumption: “the more often two articles are co-cited in document, the more likely they are related” taking proximity into account.
- Initial evaluation on 350k papers and 1,200 human relevance judgements shows a ~25% increase in F_1 score



Publications on this work

- Knoth, P., Anastasiou, L., Charalampous, A., Cancellieri, M., Pearce, S., Pontika, N. and Bayer, V. (2017)

Towards effective research recommender systems for repositories

, Open Repositories 2017, Brisbane, Australia

- Knoth, P. and Khadka, A. (2017)

Can we do better than co-citations? Bringing Citation Proximity Analysis from idea to practice in research articles recommendation

, 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries, @SIGIR 2017, Tokyo, Japan

- Charalampous, A. and Knoth, P. (2017)

Classifying document types to enhance search and recommendations in digital libraries

TDM in Research Evaluation

A faint, light gray background network diagram consisting of numerous interconnected nodes and lines, resembling a molecular or social network, is spread across the entire slide.

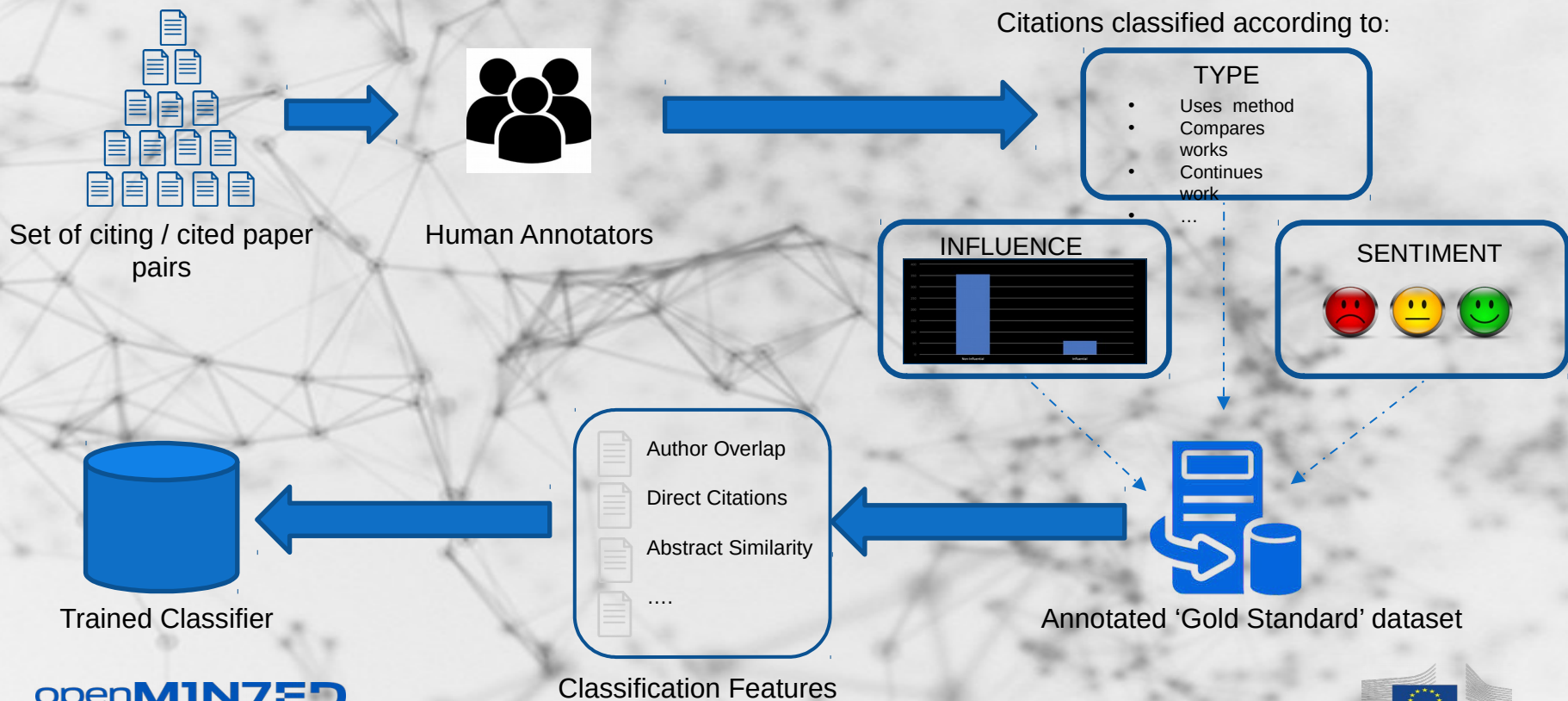
SEMANTOMETRICS

- A class of research evaluation metrics that measures research value by analysing the full texts of publications.
- Semantometrics aim to measure how far each scientific discovery takes us.
- "Reading and judging a researcher's work is much more appropriate than relying on one number." – Leiden Manifesto

TDM in citation analysis

- Current quantitative research evaluation methods are largely based on citation counts.
 - Journal Level – Journal Impact Factor (JIF)
 - Author Level – h-index, g-index
- All citations are equal, but some are more equal than others ...
- None of these metrics account for citation type or sentiment.
- Open Access means increased availability of full-text papers and articles for TDM analysis.

Detecting citation importance



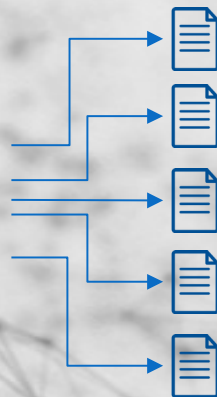
Detecting citation importance

Citing / Cited Paper Pairs

INPUT: Paper X



Citation Extraction



[1] Knoth, P., Anastasiou, L., Charalampous, A., Cancellieri, M., Pearce, S., Pontika, N., Bayer, V.: Towards effective research recommender systems for repositories. In: Proceedings of Open Repositories 2017

[3]

[4]

[n]

Paper, Citation, Label

X, [1], incidental
X, [2], incidental
X, [3], influential
X, [4], incidental
X, [n],

Classifier

Author Overlap
Direct Citations
Abstract Similarity
....

Feature Extraction

Analysis of features

- Many features used for this task by researchers, examples:
 - Total number of direct citations
 - Number of direct citations per section
 - Total number of indirect citations and number of indirect citations per section
 - Author overlap (Boolean)
 - Citation is considered helpful (Boolean)
 - Citation appears in table or caption
 - $1 / \text{Number of references}$
 - Number of paper citations / all citations
 - Similarity between abstracts
 - PageRank
 - Number of citing papers after transitive closure
 - Field of cited paper.
- Challenge: fairly small evaluation datasets

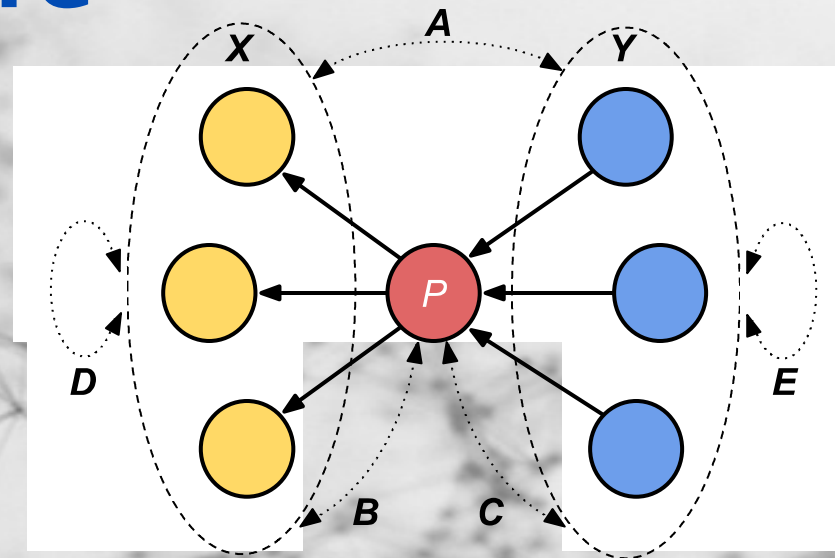
Contribution measure

Assumption: Added value of publication p can be estimated based on the semantic distance from the publications cited by p to publications citing p .



Contribution measure

- Based on semantic distance between citing and cited publications
 - Cited publications – state-of-the-art in the domain of the publication in question
 - Citing publications – areas of application
- Tested 100 different distance combinations.
- Detailed explanation and formula at semantometrics.org



True Impact Dataset (TID)

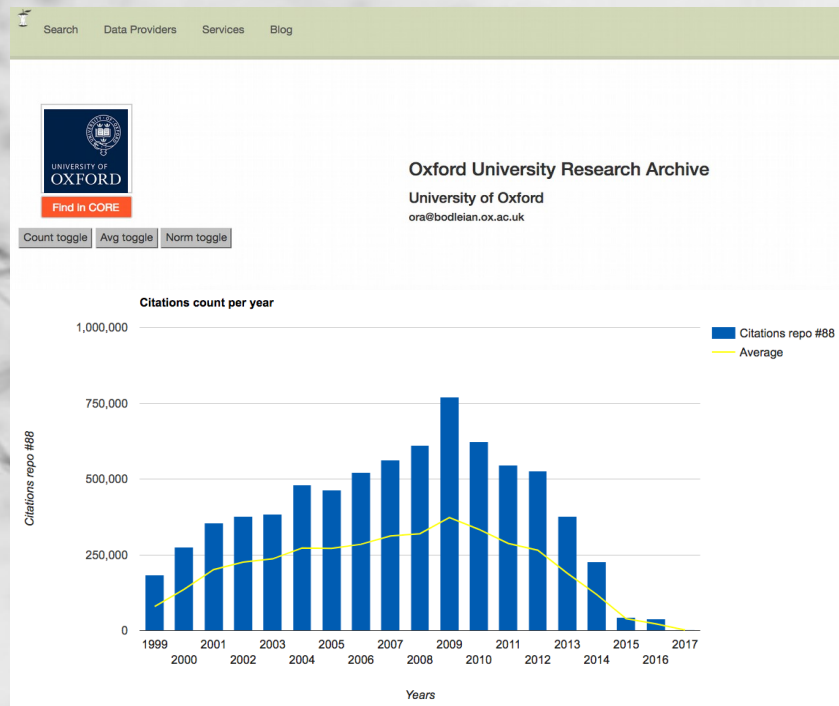
- Seminal and survey papers: two extreme cases of paper types with different type of contribution:
 - Seminal: massive contribution to knowledge generation
 - Survey: educational value, but no contribution to knowledge generation
- Key idea: A good research evaluation metric should be able to distinguish between these two publication types

True Impact Dataset (TID)

- Experimental results:
 - Citation counts (~60% accuracy, i.e. 10% over baseline)
 - Readership (does not perform better than baseline)
 - Both metrics only poorly distinguish between seminal and survey papers.
- We managed to achieve better results with the contribution method on this task than with widely used citation counts.

CORE Research Analytics Dashboard

- A prototype service for universities helping them to track research impact.
- TDM to slice and dice the data by department, funder and field
- Benchmarking metrics against others
- Integration of semantometrics in the future



Publications on this work

- **Herrmannova, D., Patton, R., Knoth, P. and Stahl, C. (2017) Citations and readership are poor indicators of research excellence: Introducing TrueID, a new dataset for validating research evaluation metrics**, Workshop: Scholarly Web Mining (SWM) at Tenth ACM International Conference on Web Search and Data Mining (WSDM2017)
- **Pride, D. and Knoth, P. (2017)**
[Incidental or influential? A decade of using text-mining for citation function classification](#)
, 16th International Conference on Scientometrics & Informetrics, Wuhan, China
- **Pride, D. and Knoth, P. (2017)**
[Incidental or influential? - Challenges in automatically detecting citation importance using publication full texts](#)
, 21st International Conference on Theory and Practise of Digital Libraries (TPDL), Thessaloniki, Greece
- **Knoth, P. and Herrmannova, D. (2014)**
[Towards Semantometrics: A New Semantic Similarity Based Measure](#)

Contributions

- Two OpenMinTeD applications we have built in the scholarly communications use case.
- TDM components are needed in both recommender systems and research evaluation.
- Ongoing research in both areas
- OpenMinTeD simplifies building such applications.