



Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

Real-World Recommender Systems for Academia: The Pain and Gain in Building, Operating, and Researching them

BIR2017: 5th International Workshop on Bibliometric-enhanced Information Retrieval (Keynote)

Dr. Joeran Beel

beelj@tcd.ie

Assistant Professor in Intelligent Systems, Trinity College Dublin, Ireland

2017-04-09

Outline

- 1. Introduction**
- 2. My recommender systems**
- 3. Experiences with bibliometrics**
- 4. A few more things**

1. Introduction

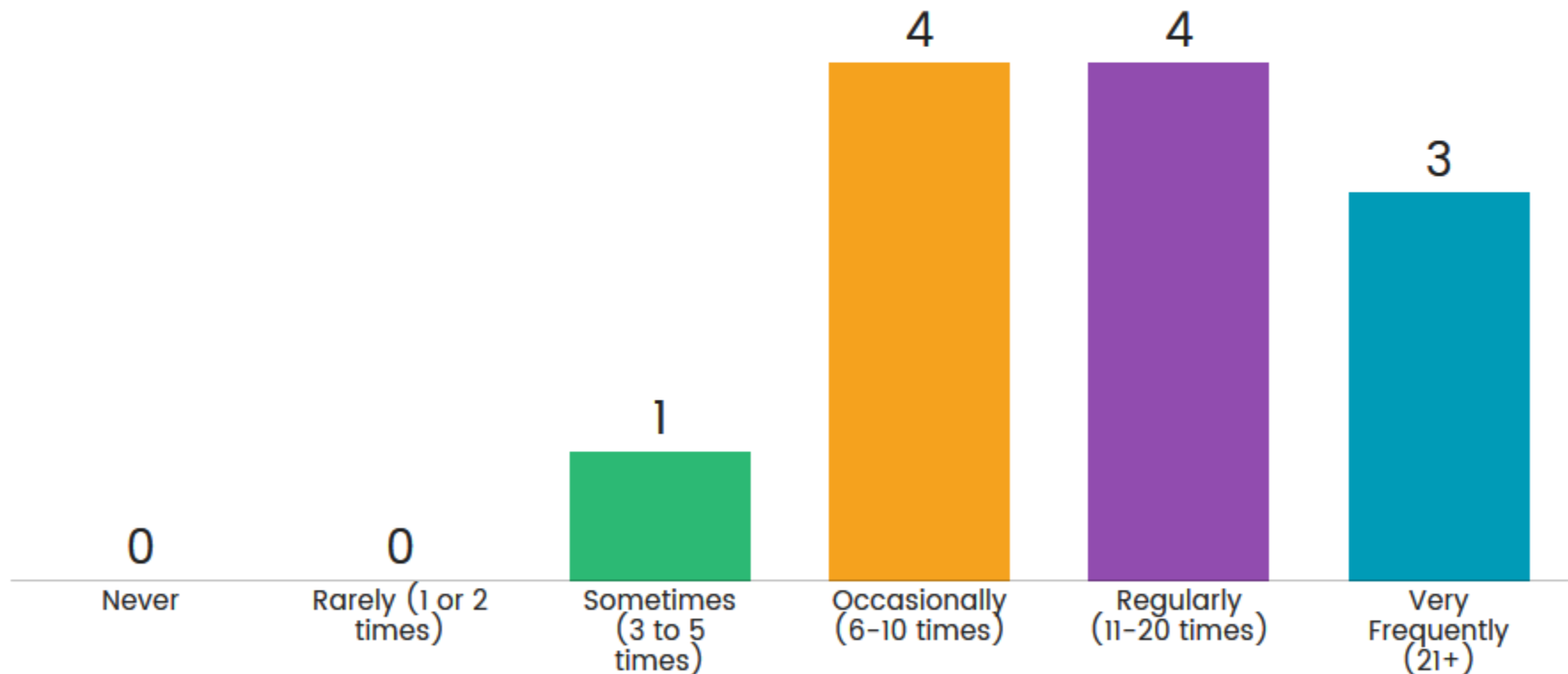
Recommender Systems in General

The image is a collage of four screenshots from different platforms, each highlighting a specific recommendation system:

- Spotify (Top Left):** Shows the 'Discover Weekly' playlist. The title 'Discover Weekly' is prominently displayed. Below it, a description reads: 'Your weekly mixtape of fresh music. Enjoy new discoveries and deep chosen just for you. Updated every Monday, so save your favourites! Created by: Spotify • 30 songs, 1 hr 40 min'. There are 'PAUSE' and 'FOLLOWING' buttons.
- Amazon (Top Right):** Shows a product page for camera lenses. A red box highlights the section 'Related to items you've viewed' with a 'See more' link. Below this, three camera lenses are displayed.
- Amazon (Middle Right):** Shows a section titled 'Inspired by your shopping trends'. A red box highlights this title. Below it, several books are displayed, including 'ELECTRONIC COMMERCE' by Gary P. Schneider, 'BUSINESS ETHICS', 'THE BUSINESS OF FASHION', 'Systems Analysis and Design Tenth Edition', and 'PROMO'.
- Amazon (Bottom Right):** Shows a section titled 'Recommendations for you in Health & Household'. A red box highlights this title. Below it, several health and household products are displayed, including '5HTH-EDGE', 'CREATINE DNA', 'TWINLAB RIPPED FUEL', 'BASCHA ISOLATE', 'platinum creatine', and 'Dymatize'.
- Netflix (Bottom Left):** Shows the Netflix homepage. The 'Popular on Netflix' section includes 'THE PEOPLE O.J. SIMPSON', 'Kingsman THE SECRET SERVICE', and 'Hitler A CAREER'. The 'Recently Added' section includes 'LIAM NEESON TAKEN 3', '12 MONKEYS', 'TOM HANKS CAPTAIN PHILLIPS', and 'FALLING SKIES'.

How often have you used recommender systems in an non-academic context (e.g. on Amazon, Netflix, Spotify, ...) in the past year?

 Mentimeter



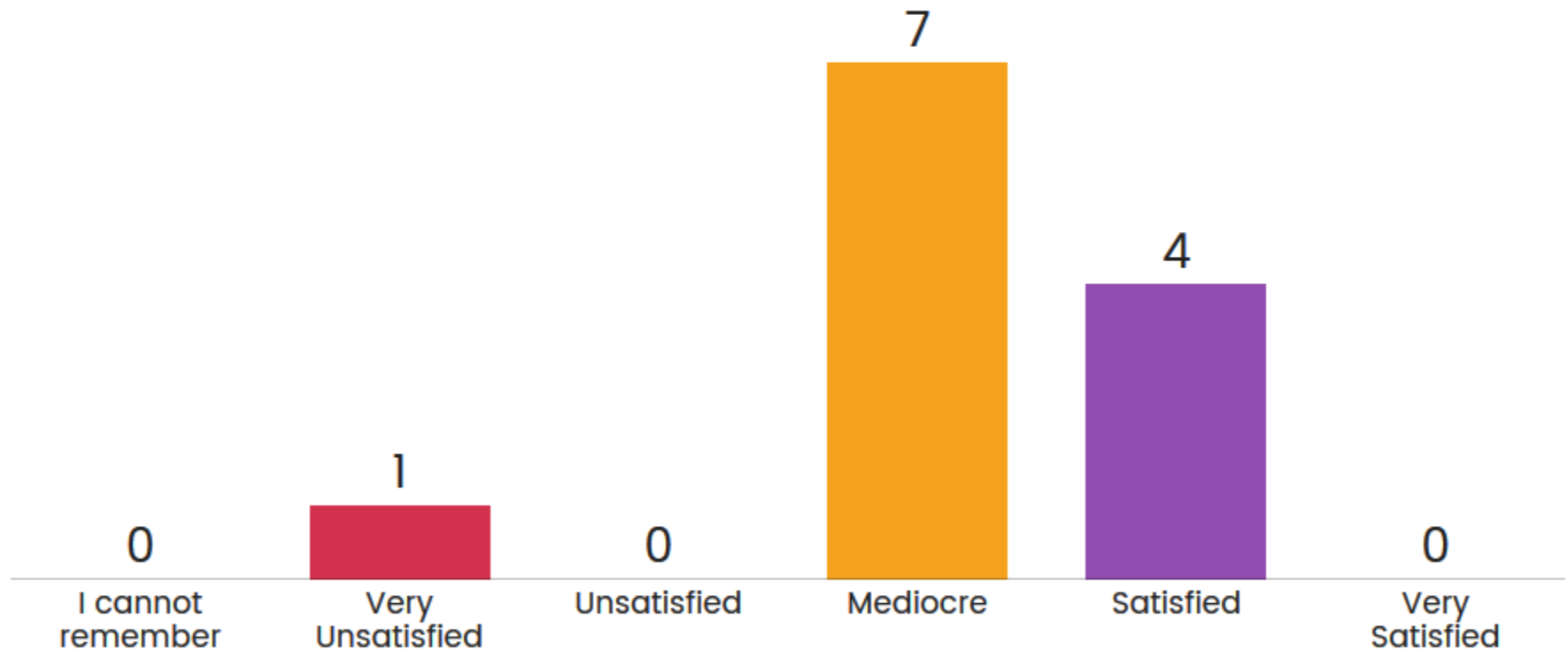
This question is not active

Activate question

 12

How satisfied were you with the recommendations?

Mentimeter



This question is not active

Activate question

12

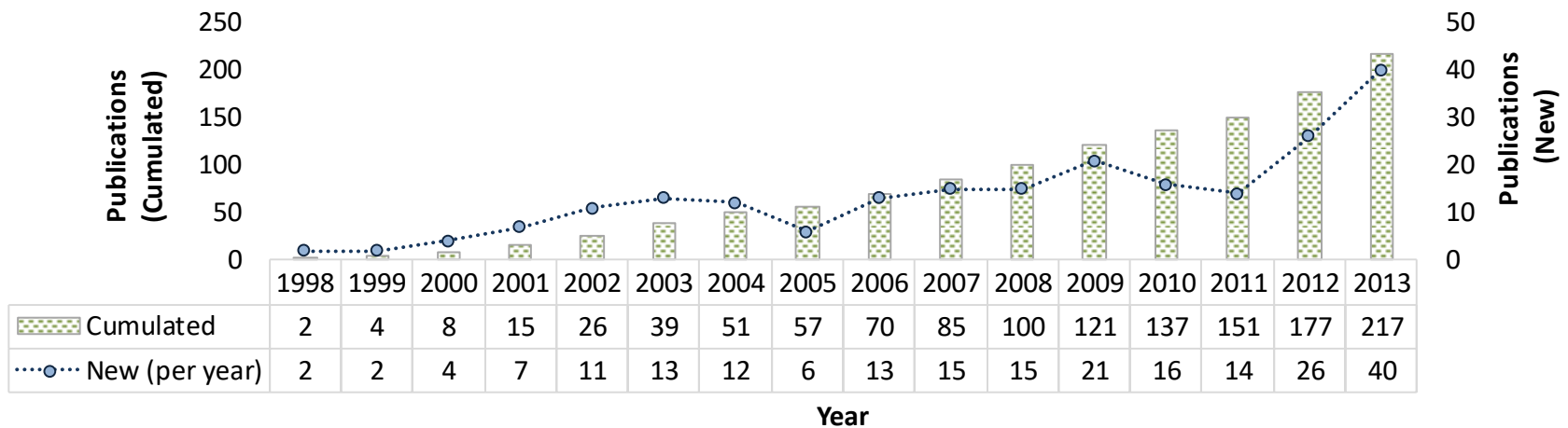
Recommender Systems in Academia

Google Scholar search results for "Increasing serendipity of recommender system with..." and "articles on Mendeley". The search shows 39 results in 0.05 seconds. The top result is "Evaluating a threefold intervention framework for assisting researchers in literature review and manuscript preparatory tasks" by A Sesagiri Raamkumar, S Foo... (Journal of ..., 2017 - emeraldinsight.com). The second result is "A knowledge sharing approach for R & D project team formation" by SM Hosseini, P Akhavan... (VINE Journal of Information ..., 2017 - emeraldinsight.com). The third result is "A personalised movie recom..." by V Subramaniaswamy, R Logesh... (13 days ago - Over the last decade, t... media, e-commerce and overall digitis... informed choices, predict marketplace... Import into BibTeX Save More).

Gmail inbox showing an email from Mendeley titled "Increasing serendipity of recommender system with..." and "articles on Mendeley". The email is dated Mar 24 and includes a "Reply" button. The email content shows a personalized suggestion for articles to read based on the user's Mendeley library.

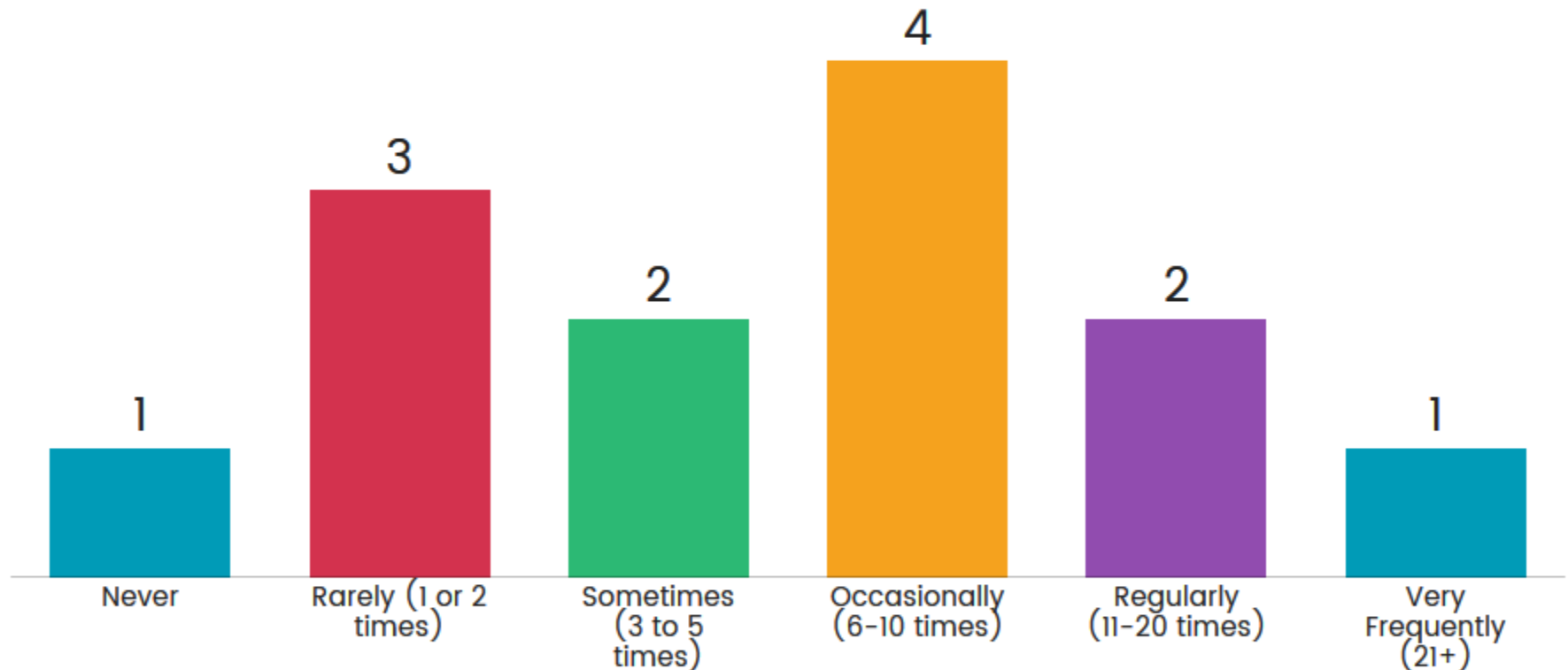
PubMed article titled "Modeling when people quit: Bayesian censored geometric models with hierarchical and latent-mixture extensions." by Okada K¹, Vandekerckhove J², Lee MD³. The abstract discusses modeling the distribution of how many items people collect before they quit, involving untagging these two possibilities. The article is from Behav Res Methods, 2017 Mar 31; doi: 10.3758/s13428-017-0879-5. [Epub ahead of print].

Increasing number of research



J. Beel, B. Gipp, S. Langer, and C. Breiteringer, "Research Paper Recommender Systems: A Literature Survey," *International Journal on Digital Libraries*, no. 4, pp. 305–338, 2015.

How often have you used recommender systems in an academic context (e.g. on Mendeley, Researchgate, PubMed, ...) in the past year?



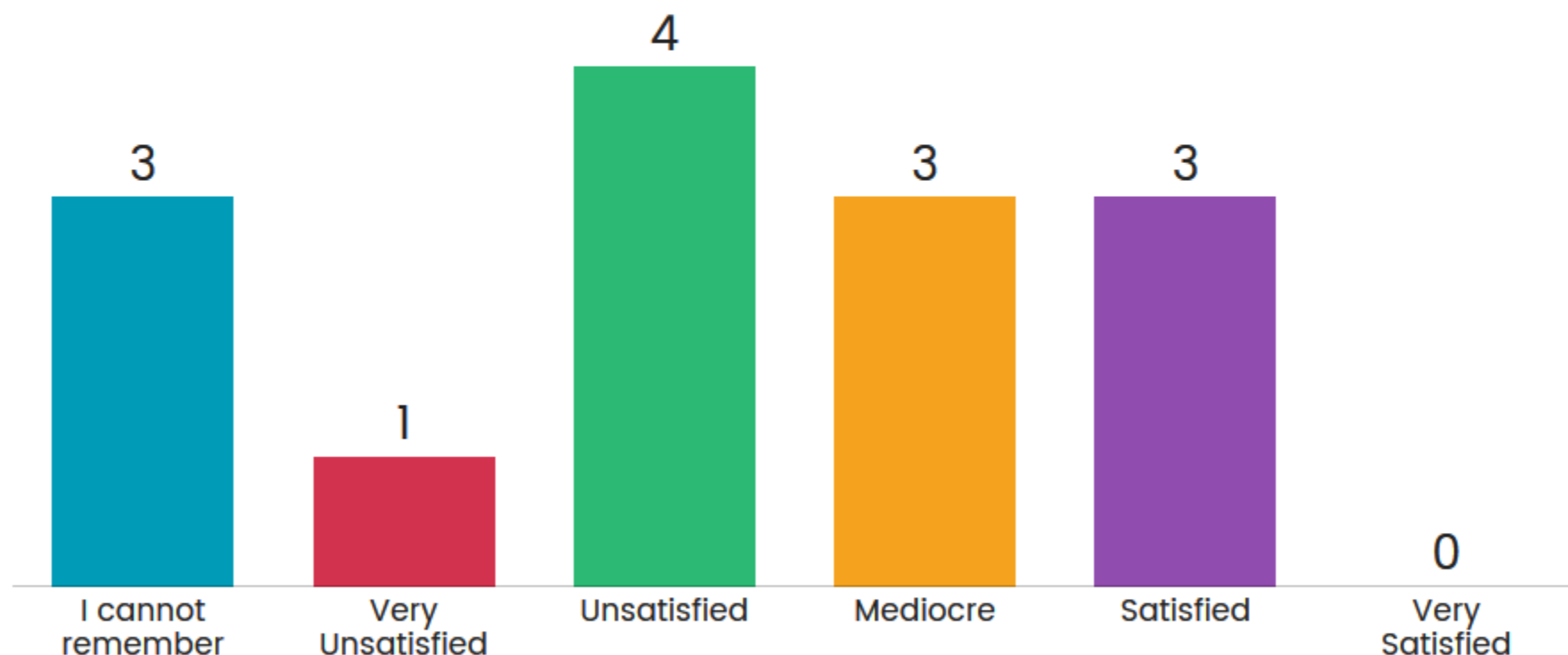
This question is not active

Activate question

 13

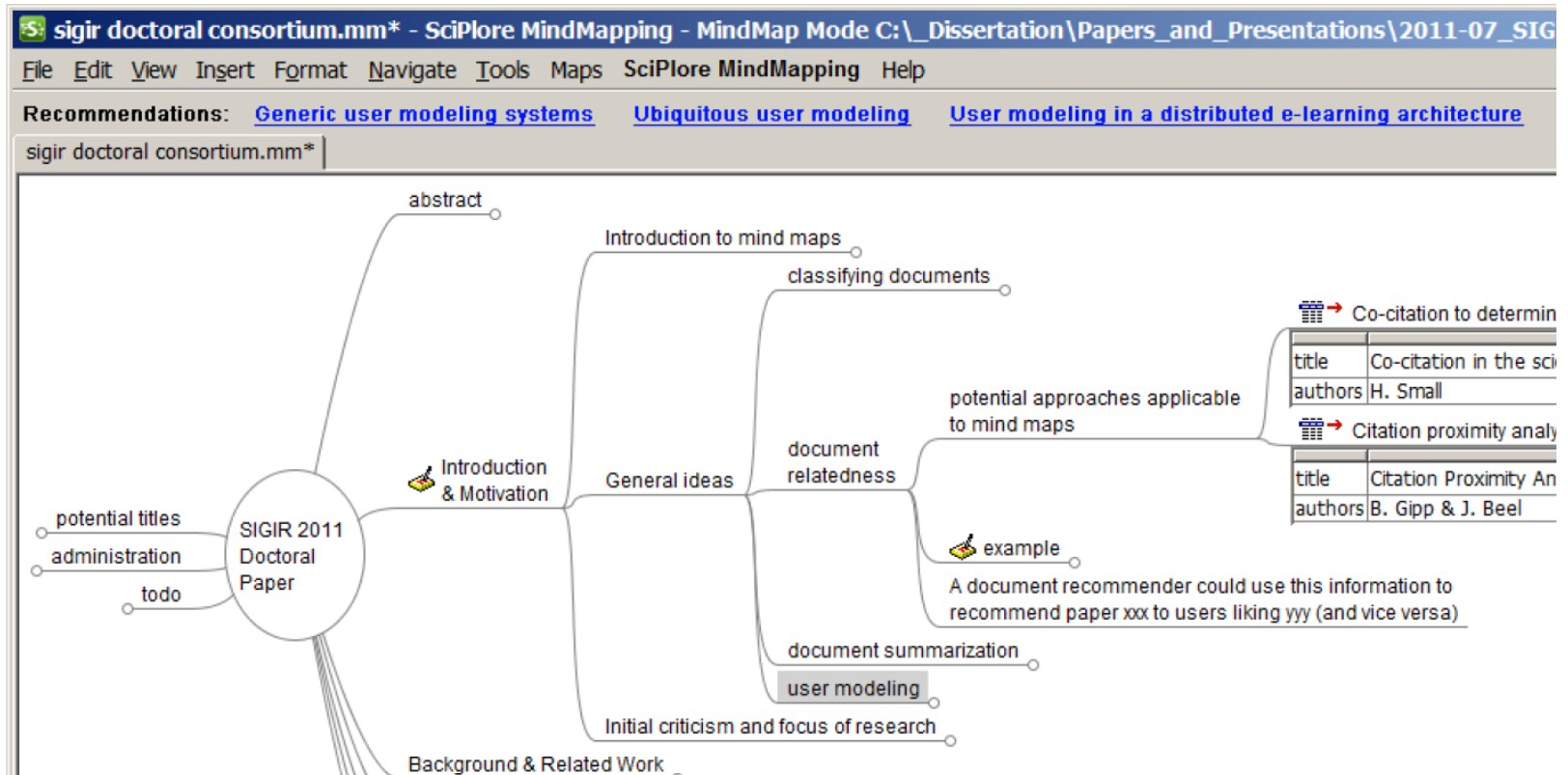
How satisfied were you with the recommendations?

Mentimeter



2. My Recommender Systems

SciPlore MindMapping



Docear

<http://docear.org>

literature_and_annotations.mm<2> - Docear - Mind map mode C:\Dissertation\Docear\Projects\Demos\Docear Teaser Docear 1.0_data\140A0BF62B2BAUL00TKZ7BUDXNBBWT00KXL2\default_files\literature_and_ann...

Home Nodes Project Files and monitoring References Resources Formatting Search and filter Navigate Tools and settings View Help

161% Zoom Minimize ribbon Center selected node Outline view Maximize map view Full screen mode Presentation mode View mode

Attribute options Display note panel Notes Hide node details Note panel position Elements

Display tool tips Display node styles in tool tips Display modification times Toolbars Miscellaneous

Library

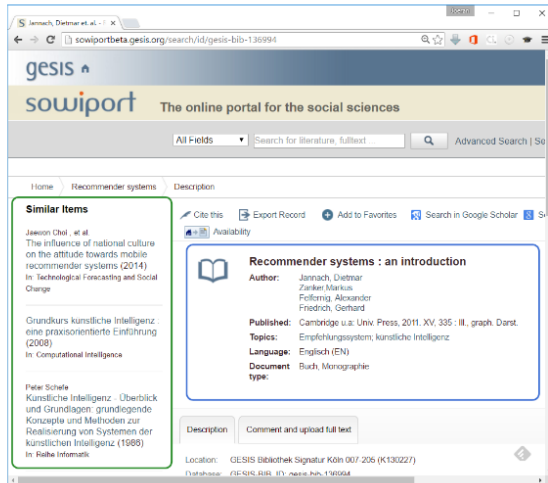
- (Academic) Search Engine Spam**
 - Detecting search engine spam from a traceback network in blogspace.pdf
 - On the robustness of Google Scholar against spam.pdf
 - Google Scholar is far easier to spam than the classic Google Search for web pages
 - Google Scholar applies no or only very rudimentary mechanisms to detect and prevent spam
 - Detecting spam web pages through content analysis.pdf
 - Academic search engine spam and Google Scholars resilience against it.pdf
 - Google Scholar indexed invisible text
 - It was easy to manipulate citation counts on Google Scholar
- Academic Search Engines**
 - Google Scholar**
 - 4,530 PowerPoint presentations, and 397,000 MS Word Documents are indexed by Google Scholar
 - Scopus
 - PubMed
- Search Engine Optimization**

journal	Journal of Electronic Publishing
authors	Beel, Joeran and Gipp, Bela
title	Academic search engine spam and Google Scholar's resilience against it
year	2010
key	Beel2010

ABC Map Version: docear 1.0 Annotation Type: COMMENT Page: 16 ObjectID: 8988505552503035580 project://140A0BF62B2BAUL00TKZ7BUDXNBBWT00KXL2/literature_repository/Academic%20search%20engine%20spam%20and%20Go

Mr. DLib

<http://mr-dlib.org>



(1) Sowipor requests related articles from Mr. DLib

Mr. DLib
Machine-readable Digital Library

(3) Sowipor displays the related articles on its website

```
<?xml version="1.0"?>
<related_articles suggested_label="Related Articles">
  <related_article document_id="5990228" original_document_id="csa-sa-201507387">
    <click_url>
      http://sowipor.gesis.org/search/id/csa-sa-201507387
    </click_url>
    <fallback_url>
      http://sowipor.gesis.org/search/id/csa-sa-201507387
    </fallback_url>
    <snippet format="html and css">
      <span class="mdl-title">The influence of national culture on the attitude towards mobile recommender systems</span>
      <span class="mdl-authors">Jaewon Choi, Hong Joo Lee, Farhana Sajjad, Habin Lee</span>
      <span class="mdl-journal">Technological Forecasting and Social Change</span>
      <span class="mdl-volume_and_number">6:66</span>
    </snippet>
    <suggested_rank>1</suggested_rank>
  </related_article>
  <related_article document_id="6080832" original_document_id="gesis-bib-116177">...</related_article>
  <related_article document_id="6084519" original_document_id="gesis-bib-5727">...</related_article>
  <related_article document_id="60891945" original_document_id="gesis-bib-5801">...</related_article>
  <related_article document_id="6329203" original_document_id="ubk-opac-HL001770873">...</related_article>
  <related_article document_id="6421430" original_document_id="ubk-opac-HL002067531">...</related_article>
  <related_article document_id="6583430" original_document_id="ubk-opac-LY000310238">...</related_article>
  <related_article document_id="6467571" original_document_id="ubk-opac-HL003359550">...</related_article>
  <related_article document_id="6075992" original_document_id="ubk-opac-LY000050421">...</related_article>
  <related_article document_id="6553806" original_document_id="ubk-opac-LY000539558">...</related_article>
  <status_reports>...</status_reports>
</related_articles>
</mr-dlib>
```

(2) Mr. DLib returns a list of related articles to Sowipor

Recommendations on Sowiport

gis
sowiport The online portal for the social sciences

All Fields Search for literature Search | Search History | Favorites

Home Search result MODERNIZATION AND... Description

Similar Items

Jack Amariglio , et al.
Postmodernism, Marxism, and the Critique of Modern Economic Thought (1994)
In: Rethinking Marxism

Michael Hout
Modernization and Postmodernization: Cultural, Economic, and Political Change in Forty-Three Societies (1998)
In: Contemporary Sociology

Amparo Coscolla , et al.
Theoretical Orientations of Spanish Psychotherapists: Integration and Eclecticism as Modern and Postmodern Cultural Trends. (2006)
In: Journal of Psychotherapy Integration

Cite this Export Record Add to Favorites Search in Google Scholar Search in Google Books

Availability #4 of 63 Next »

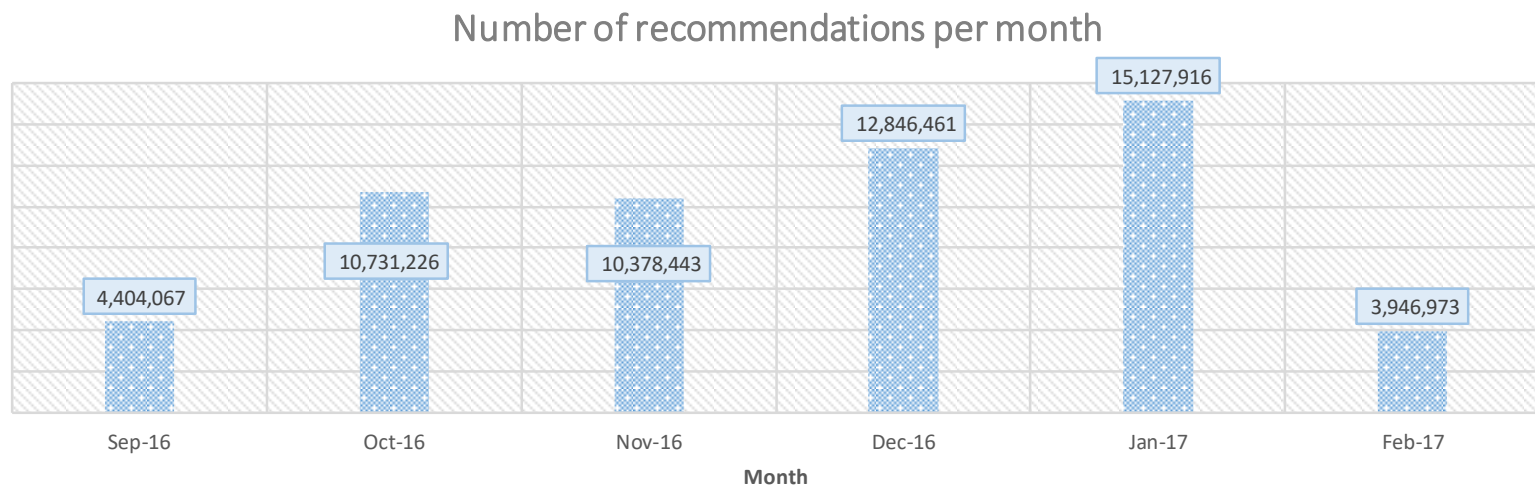
MODERNIZATION AND POSTMODERNIZATION: CULTURAL, ECONOMIC, AND POLITICAL CHANGE IN 43 SOCIETIES

Author: Inglehart, Ronald
Published: Princeton U Press 1997. ISBN 069101180X
Classification: sociology: history and theory; comparative & historical sociology
Topics: Modernization; Postmodernism; Social Change; Cultural Change; Political Change; Economic Change; Crosscultural Analysis; Postmaterialism
Language: Englisch (EN)
Document type: Buch

Description Cited by (275) Upload

EN: Examines changes in political & economic goals, religious norms, & family values & how these changes influence

Number of recommendations per month



Mr. DLib in JabRef

JabRef - C:\Dropbox\Academia\Literature\references_referenceDoear.bib (biblatex mode)

File Edit Search Groups View BibTeX Quality Tools Options Help

Search...

mypublications.bib _referenceDoear.bib

#	entrytype	author/editor	title	year	journal/booktitle	bibtexkey	ranking
30	Article	Kim et al.	Effect of Loading Symbol of Online Video on Perception of Waiting Time	2017	International Jour...	kim2017effect	
31	InProcee...	Knoth et al.	Towards effective research recommender systems for repositories	2017	Proceedings of th...	Knoth2017	
32	InProcee...	Langer and Beel	Apache Lucene as Content-Based-Filtering Recommender System: 3 ...	2017	5th International ...	Beel2017c	
33	InProcee...	Siebert et al.	Extending a Research Paper Recommendation System with Bibliomet...	2017	5th International ...	Siebert2017	
34	Procedi...	Mayr et al.	Proceedings of the Fifth Workshop on Bibliometric-enhanced Informati...	2017		Mayr2017	
35	InProcee...	Balog et al.	Overview of the trec 2016 open search track	2016	Proceedings of th...	balog2016ov...	
36	InProcee...	Baral and Li	MAPS: A Multi Aspect Personalized POI Recommender System	2016	Proceedings of th...	Baral2016	
37	Article	Beel et al.	Towards Reproducibility in Recommender-Systems Research	2016	User Modeling an...	Beel2014e	
38	Article	Beel et al.	Research Paper Recommender Systems: A Literature Survey	2016	International Jour...	Beel2014a	

Required fields Optional fields Optional fields 2 Deprecated fields General Abstract Review **MDL Related articles** {} biblatex source

- [Search engines: information retrieval in practice](#) Bruce W. Croft, Donald Metzler, Trevor Strohman. Addison-Wesley. 2009.
- [Contextual search : a computational framework](#) Massimo Melucci. *Foundations and trends in information retrieval* ; Vol. 6, No. 4-5. 2012.
- [Click models for web search](#) Aleksandr Chuklin, Ilya Markov, Marten de Rijke. *Synthesis lectures on information concepts, retrieval, and services* ; 43. 2015.
- [Ethical issues in open adoption: implications for practice](#) Frederic G. Reamer, Deborah H. Siegel. *Families in Society*. 2007.
- [The domain-specific task in CLEF 2004: overview of the results and remarks on the assessment process](#) Michael Kluck, Peters, Carol; Clough, Paul D., *Multilingual Information Access for Text, Speech and Images: 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, Berlin: Springer. (Lecture Notes in Computer Science; vol. 3491) ISBN 3-540-27420-0*. 2006.
- [Powering search: the role of thesauri in new information environments](#) Ali Shiri. *ASIS&T monograph series*. 2012.

[What is Mr. DLib?](#)

Status: JabRef is up-to-date.

Content-Based Filtering

Similar Items

Jack Amariglio , et al.
Postmodernism, Marxism, and
the Critique of Modern Economic
Thought (1994)
In: Rethinking Marxism

Michael Hout
Modernization and
Postmodernization: Cultural,
Economic, and Political Change
in Forty-Three Societies (1998)
In: Contemporary Sociology

Amparo Coscolla , et al.
Theoretical Orientations of
Spanish Psychotherapists:
Integration and Eclecticism as
Modern and Postmodern Cultural
Trends. (2006)
In: Journal of Psychotherapy Integration

MODERNIZATION AND POSTMODERNIZATION: CULTURAL, ECONOMIC, AND POLITICAL CHANGE IN 43 SOCIETIES

Author: Inglehart, Ronald

Published: Princeton U Press 1997.
ISBN 069101180X

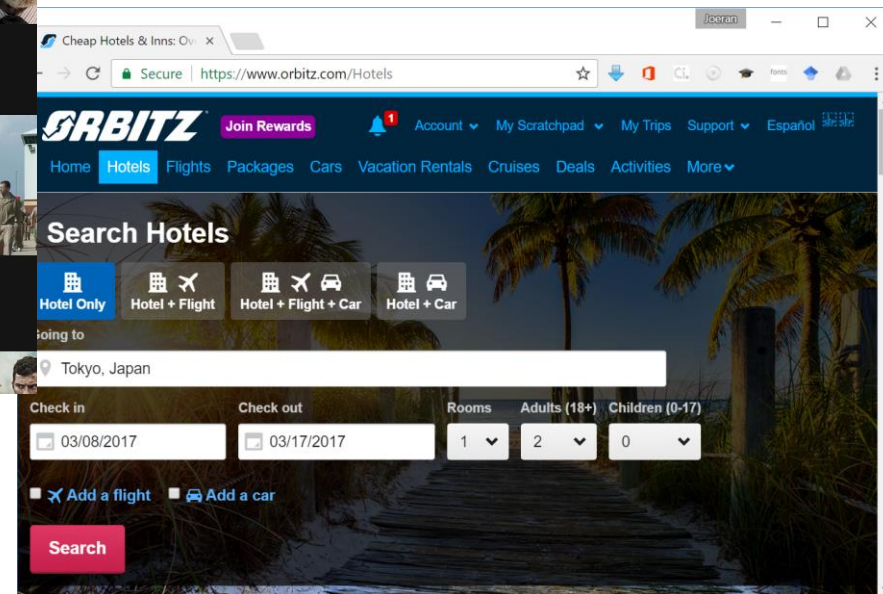
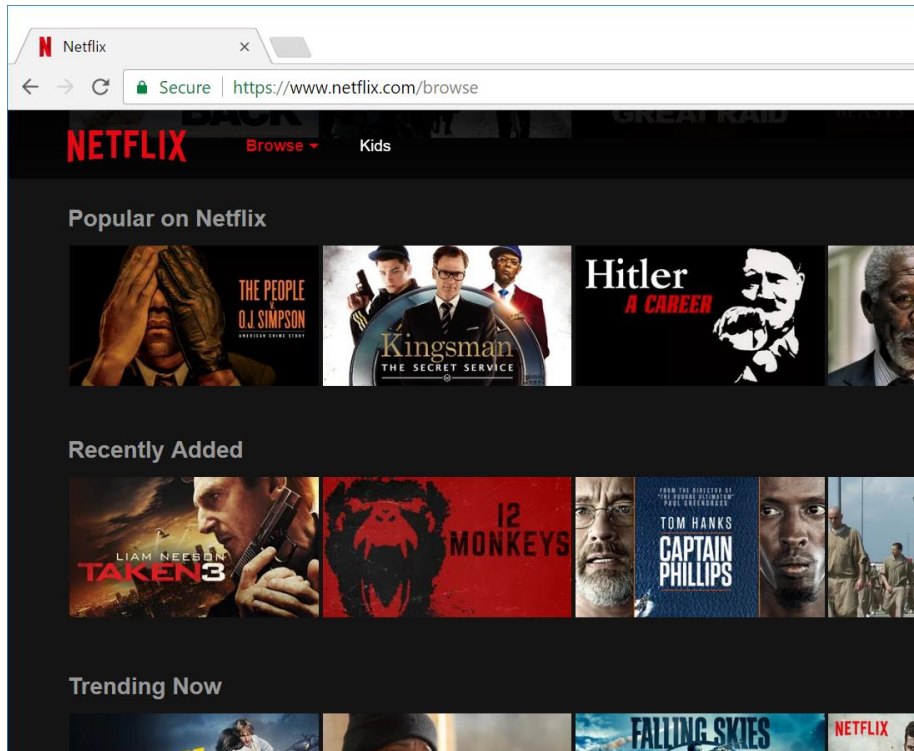
Classification: sociology: history and theory; comparative & historical sociology

Topics: Modernization; Postmodernism; Social Change; Cultural Change; Political Change; Economic Change; Crosscultural Analysis; Postmaterialism

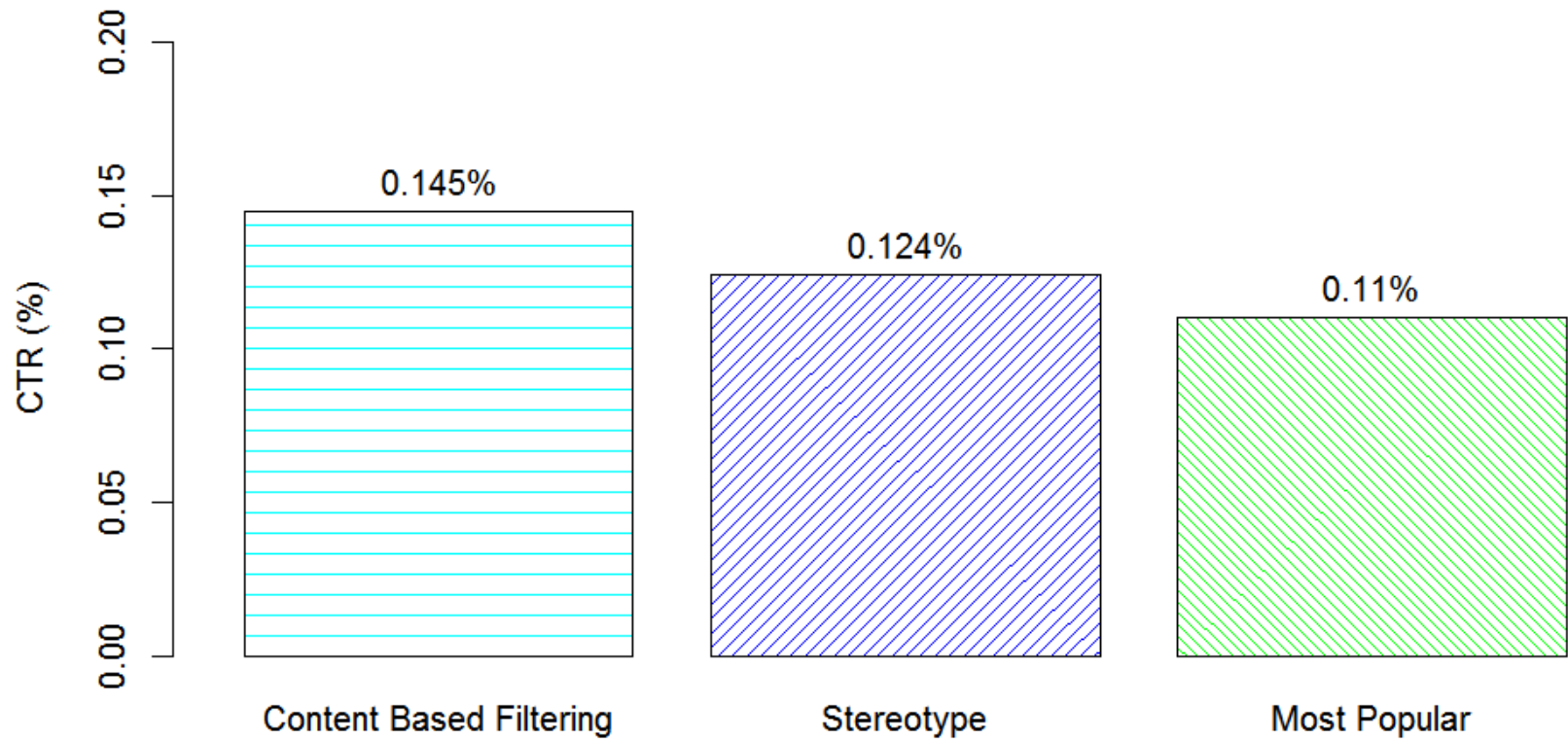
Language: Englisch (EN)

Document type: Buch

Stereotype and Most-Popular Recommendations



Overall



Required Labor

SciPlore MindMapping

- A few weeks by a student

Docear

- 18 person months over three years
- + time for „host“ application

Mr. DLib

- 2 students and 2 postdocs each 2 month full time for first prototype
- 30 FTE months since last year Mai

Netflix, Amazon, ...: much much more

Required Skills

Server Administration

Databases

Web Technologies / Web Services

Data Processing

Scalability

Data Privacy

Project Management

Niche knowledge, e.g. Bibliometrics

3. Experience with Bibliometrics

Characteristics of Content-Based Filtering

Most common approach for research-paper recommender systems

- Can be applied with almost all documents
- No cold start problem

Solely focused on content

No quality/impact assessments

Factors to evaluate relevance in Academia

- **Decision to Read**
 - Content \leftarrow Content Based Filtering
 - Title
 - Abstract
 - Quality \leftarrow Bibliometrics
 - Venue Reputation
 - Author Reputation
- **Decision to cite**
 - Content \leftarrow Content Based Filtering
 - Quality \leftarrow Bibliometrics

Getting Bibliometrics for Research Only

- **Use one of the many datasets that seem best for your research**
- **If it's not enough, take another one, ...**
- **Tweak them if necessary**
- **Do research**

Tweking

Caragea et al. removed papers with fewer than ten and more than 100 citations from the evaluation corpus, as well as papers citing fewer than 15 and more than 50 papers. From originally 1.3 million papers in the corpus, around 16,000 remained (1.2%)

Pennock et al. removed documents with fewer than 15 implicit ratings from the corpus. From originally 270,000 documents, 1,575 remained (0.58%).

In Real-World

Document corpus is given – tweaking or extending it is not really possible

In case of Sowiport

- **Mostly meta data (no full-text)**
- **Rather few references**
- **Only Sowiport documents should be recommended**

In case of CORE

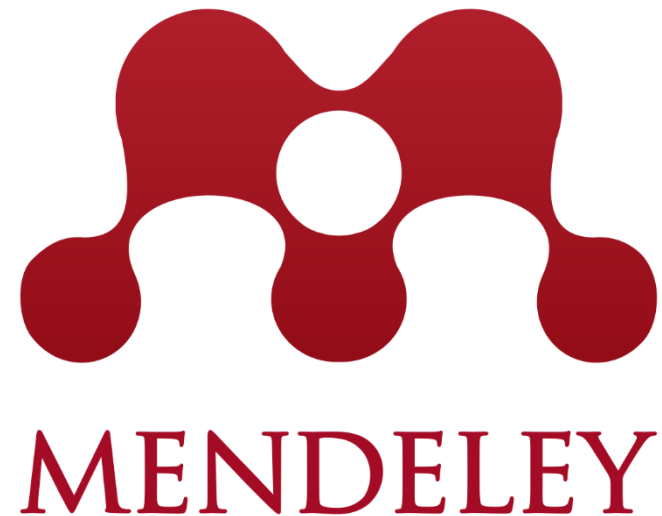
- **Full-text as PDFs**
- **Rather few references**

Readership data from Mendeley

Freely available via API

Rather good coverage (so we hoped)

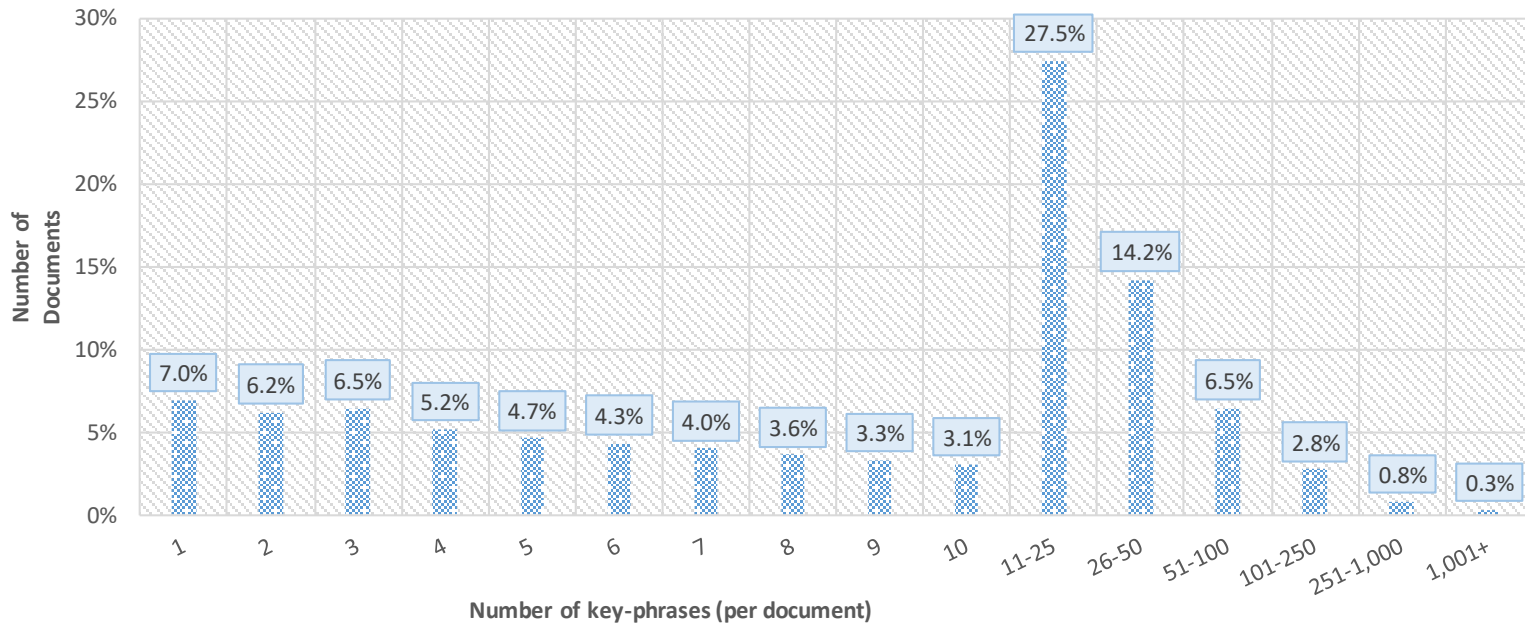
Timely



Coverage

18% of Sowiports documents have at least one reader

CORE has higher coverage



Experiment

Select top x candidates; $x = 10...100$

Re-Rank the candidates

- **Calculate bibliometric relevance with metric m ; m = absolute readership count, readership count normalized by year, readership count normalized by number of authors, h-index**
- **Combine bibliometric relevance and text relevance score:
bibliometric only; simple multiplication; bibliometric * root(text);
bibliometric * log(text)**
- **Recommend top y candidates; $y = 1...15$**

Results (Preliminary Analysis)



Data Quality

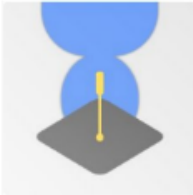
Author	Document Count
et al.	35,865
and others	26,331
AnoN.	25,719
Anonymous	21,933
[Unknown]	17,210
[[author]]???	17,191
Unknown	16,233
u.a.	16,162

et al. - Google Scholar Citations

Secure | <https://scholar.google.de/citations?>

Google Scholar

Follow



et al.

The academic superstar everybody wants to be co-author with. See Homepage for background & creator.

[bibliometrics](#), [scientometrics](#), [citation analysis](#)

No verified email - [Homepage](#)

Citation indices		All	Since 2012
Citations		2541508	739104
h-index		332	265
i10-index		332	332

Co-authors [View all...](#)

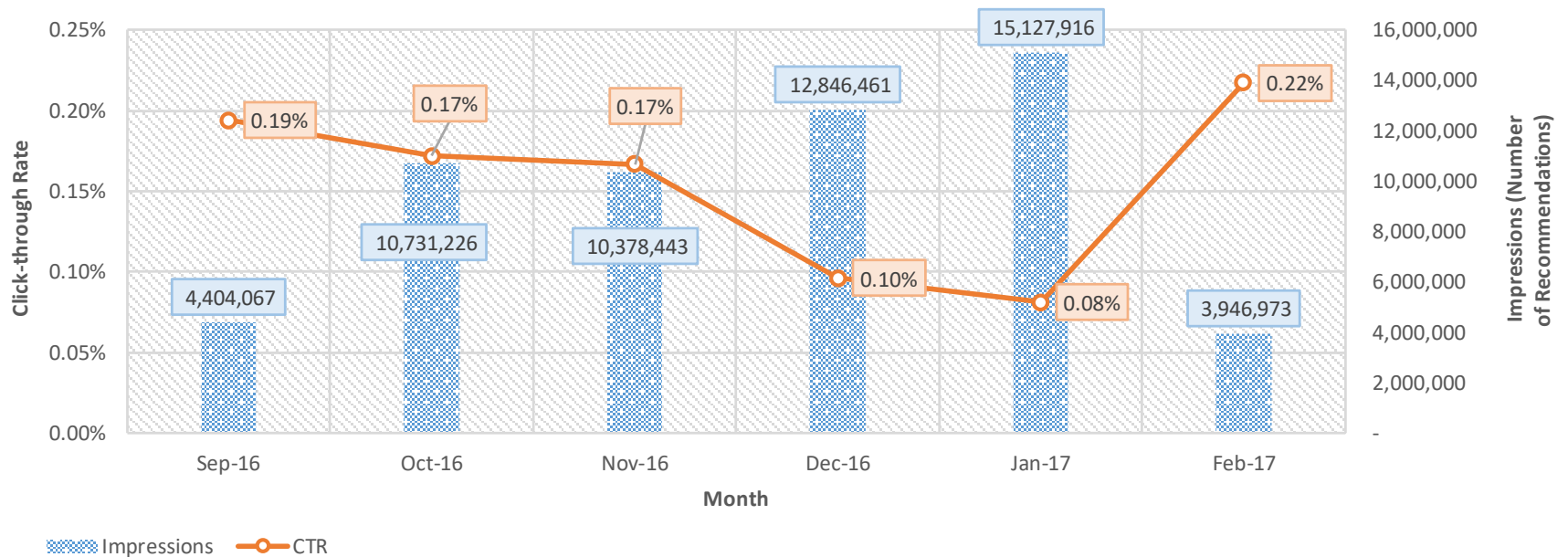
Paul Erdős, A. Author, Mark Dingemans

Title	1-20	Cited by	Year
Protein measurement with the Folin phenol reagent	OH Lowry, NJ Rosebrough, AL Farr, RJ Randall J biol Chem 193 (1), 265-275	206011	1951
Molecular cloning	J Sambrook, EF Fritsch, T Maniatis Cold spring harbor laboratory press 2, 14-9.23	175581 *	1989
Psychometric theory	JC Nunnally, IH Bernstein, JMF Berge McGraw-Hill	88037	1967

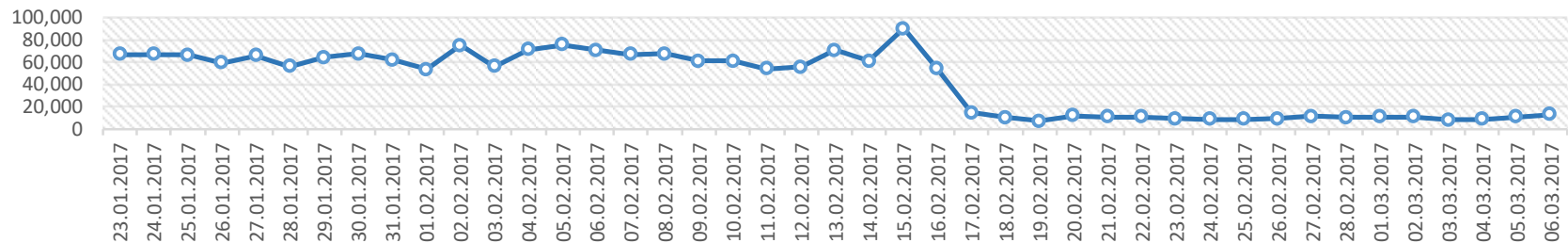
Re-rank 100 candidates?



Low Click-Through Rates / Statistical Significance



Introduction of JavaScript



CTR appropriate?

?

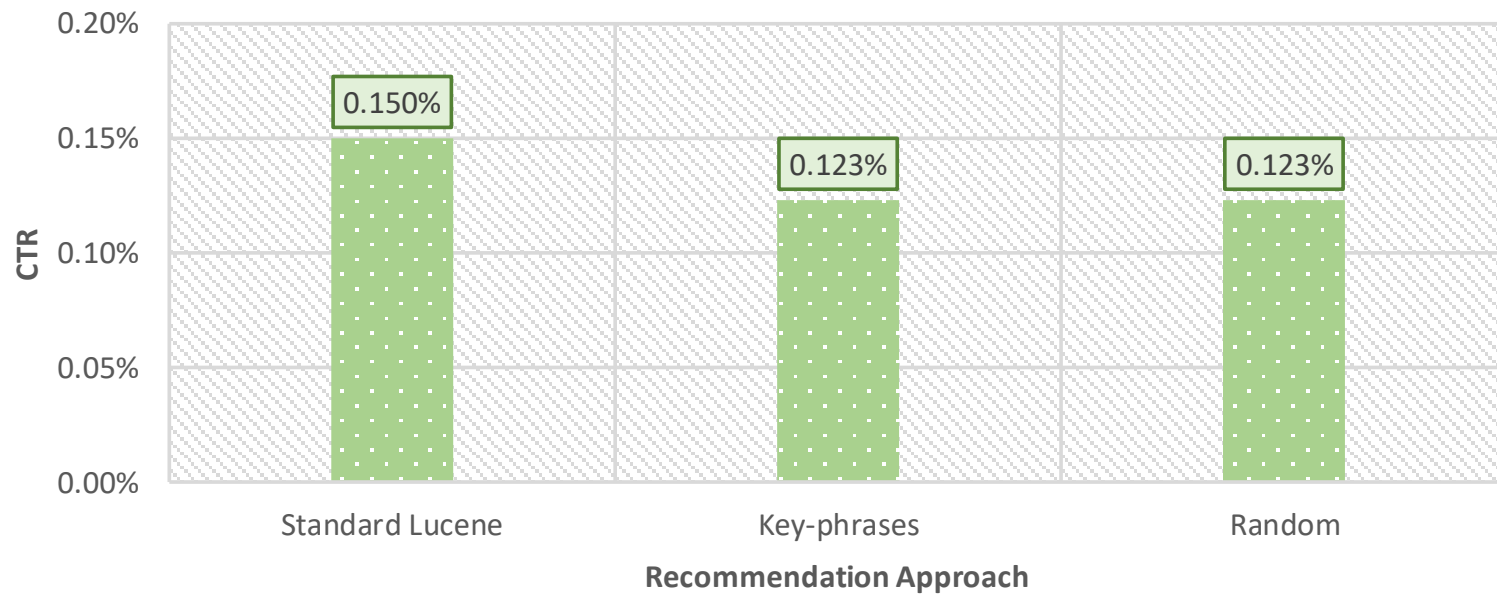


4. A few more things

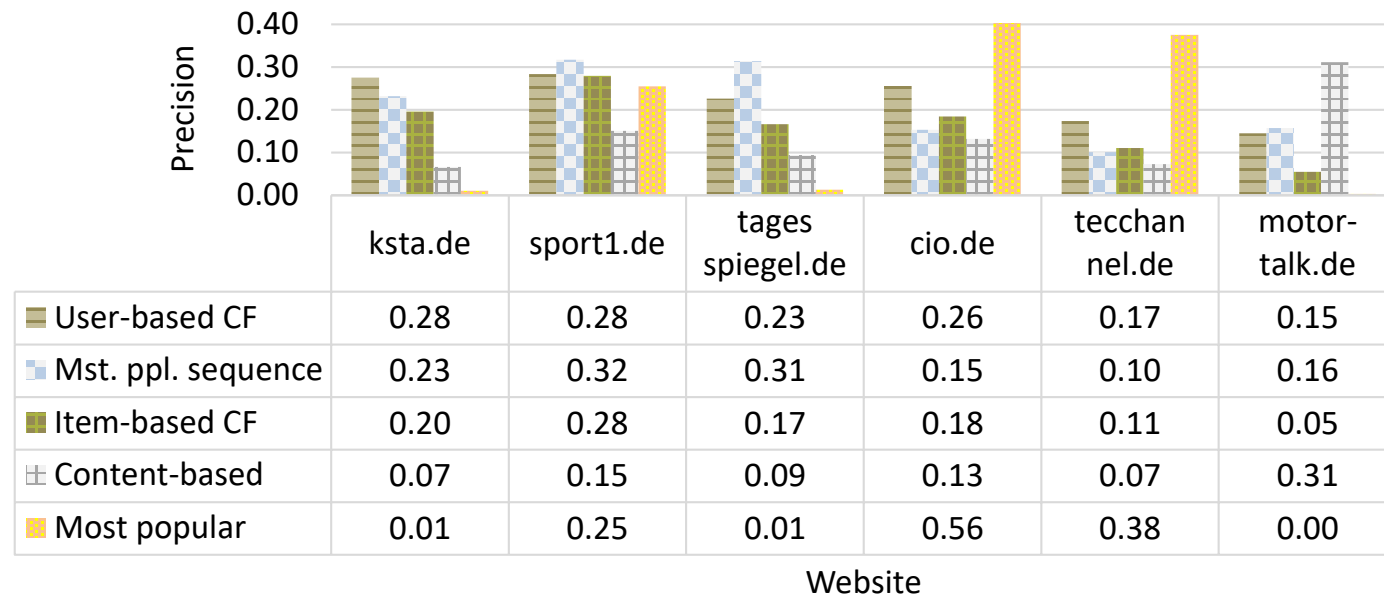
Attract students and volunteers



Get used to failure

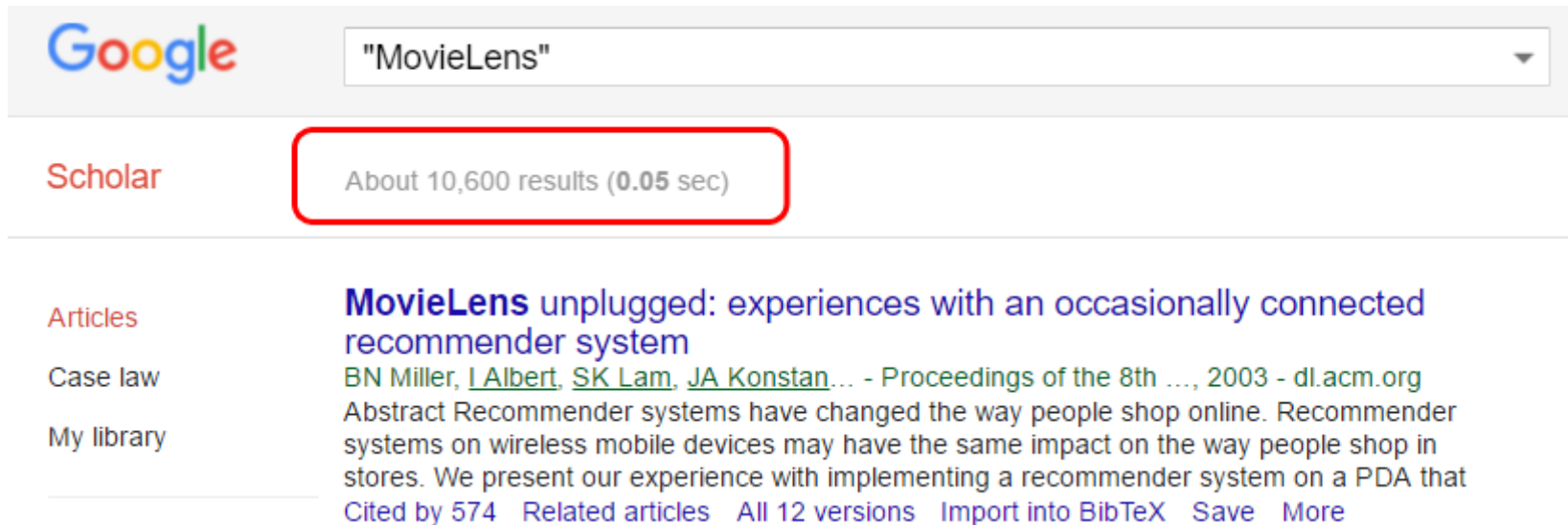


(Non) Reproducibility



Interest in datasets

MovieLens dataset was downloaded 140,000 times in 2014



The image is a screenshot of a Google Scholar search interface. At the top, the Google logo is on the left, and a search bar contains the text "MovieLens". Below the search bar, the word "Scholar" is displayed in red. To its right, a red-bordered box contains the text "About 10,600 results (0.05 sec)". On the left side of the page, there is a vertical menu with the following items: "Articles", "Case law", and "My library". The main content area displays the search results for "MovieLens". The first result is titled "MovieLens unplugged: experiences with an occasionally connected recommender system" in blue. Below the title, the authors "BN Miller, I Albert, SK Lam, JA Konstan..." are listed, followed by the publication information "- Proceedings of the 8th ..., 2003 - dl.acm.org". The abstract text reads: "Abstract Recommender systems have changed the way people shop online. Recommender systems on wireless mobile devices may have the same impact on the way people shop in stores. We present our experience with implementing a recommender system on a PDA that". At the bottom of the result, there are links: "Cited by 574", "Related articles", "All 12 versions", "Import into BibTeX", "Save", and "More".

Google "MovieLens"

Scholar About 10,600 results (0.05 sec)

Articles
Case law
My library

MovieLens unplugged: experiences with an occasionally connected recommender system
BN Miller, I Albert, SK Lam, JA Konstan... - Proceedings of the 8th ..., 2003 - dl.acm.org
Abstract Recommender systems have changed the way people shop online. Recommender systems on wireless mobile devices may have the same impact on the way people shop in stores. We present our experience with implementing a recommender system on a PDA that
Cited by 574 Related articles All 12 versions Import into BibTeX Save More

Interest in Docear dataset

31 requests to download

No Citation

No mentioning

Advertisement

1. Mr. DLib

1. Joint research on Mr. DLib (and its partners)
2. Integrate Mr. DLib in your digital library or similar
3. Integrate your content in Mr. DLib

2. Industry partnership / Science Foundation Ireland

1. Postdoctoral researcher for two years (30,000€ / year)
2. Doctoral researcher for ~ 12-15,000€ / year



Trinity College Dublin

Coláiste na Tríonóide, Baile Átha Cliath

The University of Dublin

Thank You

Joeran Beel

beelj@tcd.ie

<http://mr-dlib.org>

<http://mr-dlib.org/research/>