



Automatic detection of duplicate records in institutional repositories

Matteo Cancellieri matteo.cancellieri@open.ac.uk

Anton Zhuk anton.zhuk@open.ac.uk

Valerii Budko valerii.budko@open.ac.uk

Eka Chxaidze ekaterine.chkhaidze@open.ac.uk

Viktoriia Pavlenko viktoriia.pavlenko@open.ac.uk

Petr Knoth petr.knoth@open.ac.uk

CORE and the OA landscape



CORE's mission is

to index all open access research worldwide and deliver unrestricted access for all.

We are here to support and advance the Open Access / Open Research movement

WE ARE

the world's **most used** collection of open access research papers from repositories

WE ARE

a **not-for-profit** scholarly infrastructure dedicated to the open access mission, **adopters of POSI** principles.

WE

provide solutions for content management, discovery and scalable machine access to research.

WE

serve the global network of repositories and journals by increasing discoverability and reuse of open access content.

Joining CORE benefits your institution



Lancaster
University



University of
Strathclyde



Durham
University



White Rose
Libraries



Universities of Leeds, Sheffield & York

THE UNIVERSITY OF
CHICAGO



Imperial College
London

UNIVERSITY OF
BIRMINGHAM



University of
Nottingham

... and over 20 others
are already our
supporting and
sustaining members

Become a member to demonstrate your support for Open Research

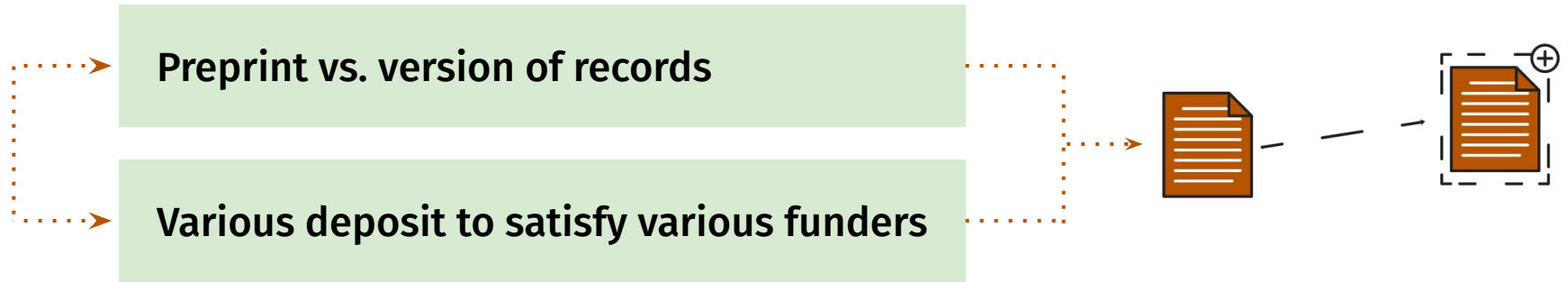
Duplicates in repositories

Is it a problem?



Duplicates in repositories

Is it a problem?

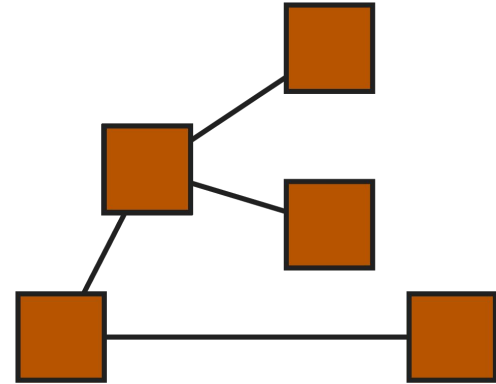


Duplicates in repositories

Is it a problem?

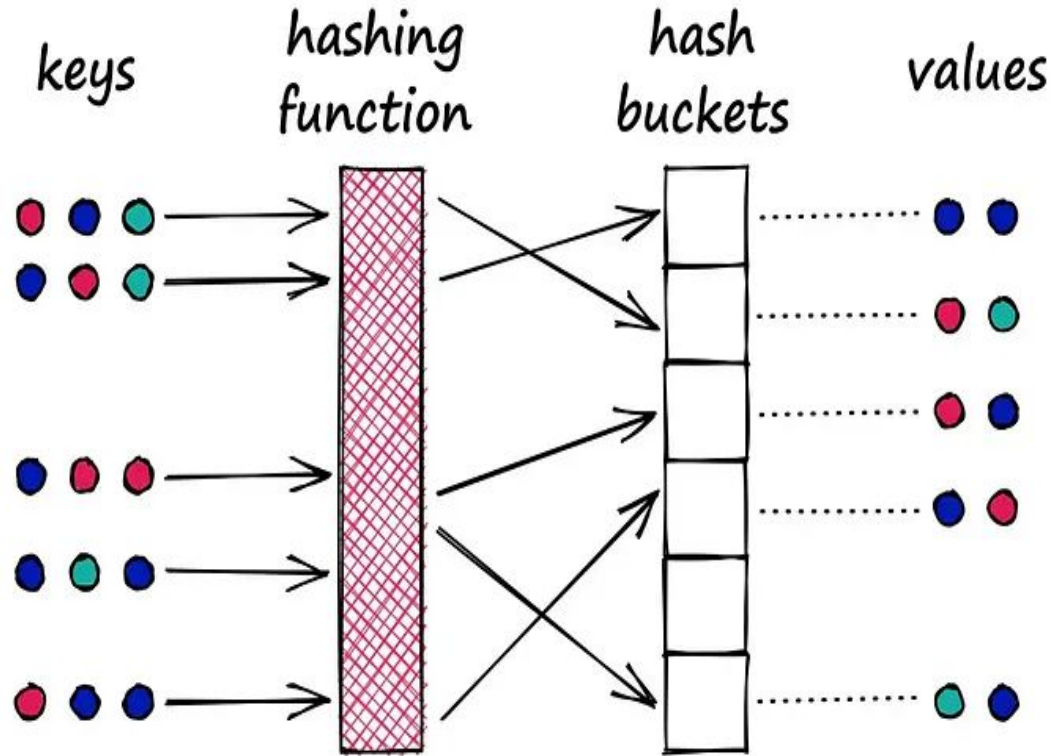


Scholarly graph degrades.



Locality sensitive hashing

Simhash to be precise



Automated detection of near-duplicates

How can you do it at scale?

simhash of A	0	0	1	0	1	0	1	1
simhash of B	0	0	1	1	1	0	1	1
$A \wedge B$	0	0	0	1	0	0	0	0

Bit population ($A \wedge B$) = 1

<https://moz.com/devblog/near-duplicate-detection>



Near duplicates in practice

Dataset from:

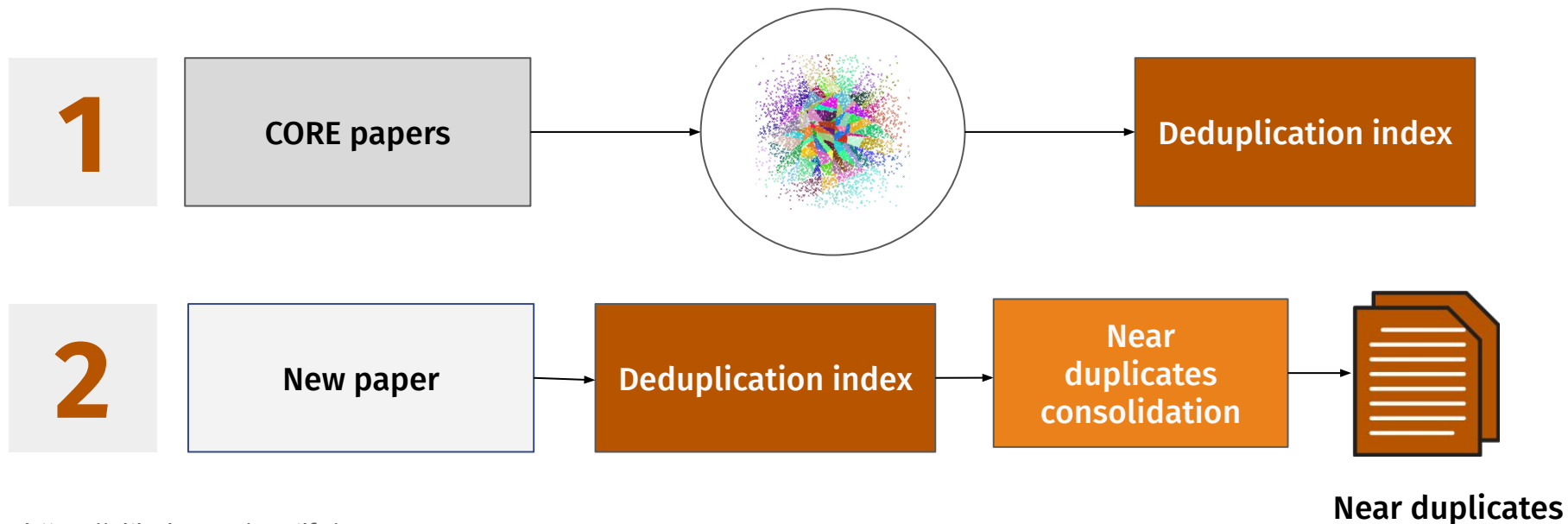
Deduplication of
Scholarly Documents
using Locality
Sensitive Hashing and
Word Embeddings

(Gyawali et al., LREC 2020)



Rules	Precision	Recall	Index duration	Search duration
Hamming distance on simhash 64 bit				
processed_title	0.83	0.77	48s	38s
processed_abstract	0.72	0.58	122s	112s
processed_title + processed_abstract	0.74	0.58	145s	131s
processed_title + processed_abstract 50	0.70	0.61	61s	53s
processed_title + processed_abstract 100	0.72	0.60	69s	60s
processed_title_initials	0.74	0.73	31s	23s
first_and_last_letters_processed_abstract + processed_title_initials	0.69	0.56	83s	75s
fullText	0.76	0.63	6201s	6162s
title+year	0.78	0.69	92s	87s
title + abstract + year	0.69	0.55	186s	178s
title + authorstring	0.58	0.55	144s	113s
title + first author	0.69	0.60	103s	87s
title + year + first author	0.68	0.57	102s	85s
Hamming distance on simhash 128 bit				
processed_title + processed_abstract	0.81	0.60	159s	154s
processed_title + processed_abstract_50	0.81	0.60	158s	153s
processed_title + processed_abstract_100	0.81	0.60	161s	147s

Large scale near duplicate checking in CORE



<https://github.com/spotify/annoy>

Finding version and near-duplicates

Versions/duplicates

List of possible duplicates

author	N/A	N/A
type	Bandara A., Charalambides M., Dulay N., Flegkas P... SHOW MORE	Bandara, Arosha K., Lupu, Emil C., Russo, Alessand... SHOW MORE
Field of study	N/A	N/A
DOI	N/A	N/A
Publication date	10.1109/tns.2006.4798308	10.1109/TNSM.2006.4798308
Deposited date	2005-05-01T00:00:00	2006-04-01T01:00:00+01:00
Abstract	2006-04-01T00:00:00	2006-04-01T00:00:00+01:00
	Policy-based management provides the ability to dynamically re-configure DiffServ networks such that desired Quality of Service (QoS) goals are achiev... SHOW MORE	Policy-based management provides the ability to dynamically re-configure DiffServ networks such that desired Quality of Service (QoS) goals are achiev... SHOW MORE

[Duplicate](#) [Different version](#) [Not the same article](#)

Comparison mode

Find versions and near-duplicates BETA

Our technology searches your repository to identify different versions of your articles and potential duplicates within your repository. This can help you in managing and curating your repository content. We periodically detect near-duplicate records and versions and allow you to compare them side by side. This can help... [SHOW MORE](#)

Last successful check

29/05/2023

Detection of versions and near-duplicates runs automatically every time after your repository is harvested.

Number of matches

Please be aware that it is a list of possible duplicates, and you can review it.

964

[DOWNLOAD](#)

List of potential duplicates and alternative versions

Search

OAI	Title	Authors	Duplicates	Publication date		
6864	Policy refinement for Di...	Bandara, Arosha K. ...	+ 2 found	2005-05-01		⋮
18692	A psychoacoustical inve...	Whitehouse, James ...	+ 2 found	2008-04		⋮

[DOWNLOAD CSV](#)

Showing 1-20 records of 466 records

[SHOW MORE](#)

Finding version and near-duplicates

Find versions and near-duplicates BETA

Our technology searches your repository to identify different versions of your articles and potential duplicates within your repository. This can help you in managing and curating your repository content. We periodically detect near-duplicate records and versions and allow you to compare them side by side. This can help... [SHOW MORE](#)

Last successful check

25/05/2024

Detection of versions and near-duplicates runs automatically every time after your repository is harvested.

Number of matches !

These matches might include potential duplicates, near-duplicates and different versions of papers and metadata records.

999

[DOWNLOAD](#)

List of potential duplicates and alternative versions ! ↓

We have found **999** items. Review and download them below.

[Show tips](#) !

i The data on this list is cached for performance reasons. The processing will take some time but you will be able to see the updated data

Search

OAI	Title	Authors	Matches	Status ⇅	Version	Publication date ⋮
95949	Housing Policy: An Introduction	Balchin, Paul Rhoden, Maureen	+ 2 found	To review	- -	2002
6385	A "Learning Revolution"? Investigating Pedago...	Gillen, Julia Kleine Staarman, Judit...	+ 2 found	Reviewed	AO SMUR	2006-04
9165	A magyar forradalom új megközelítésben: az ip...	Pittaway, Mark	+ 2 found	Reviewed	NA NA	2006-02
12574	Transverse Aeolian Ridges (TARs)on Mars	Balme, M. R Berman, D. C. Bourke,...	+ 2 found	To review	- -	2008

Finding version and near-duplicates

Metadata title

The reference paper

The record that the reference paper is compared with:
OAI 12574

The record that the reference paper is compared with:
OAI 12057

Title

Transverse Aeolian Ridges (TARs) on Mars

Transverse Aeolian Ridges (TARs) on Mars

Transverse Aeolian Ridges (TARs) on Mars

Live In CORE

Live In CORE

Open In The Repository

Live In CORE

Open In The Repository

Author

Balme, M. R., Berman, D. C., Bourke, M. C., Ralfk...
SHOW MORE

Balme, M. R., Berman, D. C., Bourke, M. C., Ralfk...
SHOW MORE

Balme, Matthew, Berman, Daniel, Bourke, Mary, Zimb...
SHOW MORE

Type

research

research

Not available

Field of study

Not available

Not available

Not available

DOI

10.1016/j.geomorph.2008.03.011

10.1016/j.geomorph.2008.03.011

10.1016/j.geomorph.2008.03.011

Publication date

2008-11-01T00:00:00

2008-11-01T00:00:00+00:00

2008-11-01T00:00:00+00:00

Deposited date

2008-10-20T05:48:00

2008-11-01T00:00:00+00:00

2008-11-01T00:00:00+00:00

Abstract

Abstract not available

Abstract not available

Aeolian processes are probably the dominant ongoing surface process on Mars; Large Dark Dunes (LDDs), particularly common aeolian landforms, were firs...
SHOW MORE

Not available for works

Please indicate the version of articles:

Please indicate the version of articles:

Version

Not available for works

AO = Author's Original

SMUR = Submitted Manuscript Under Review

AM = Accepted Manuscript

P = Proof

VoR = Version of Record

CvOR = Corrected Version of Record

EvOR = Enhanced Version of Record

NA = Not Applicable (or Unknown)

Different version (N/A)

Mark this paper as

Mark this paper as



University of Chicago adopts CORE's article deduplication tool

We do not have a duplicate check built into our repository, so CORE's duplicate check is immensely helpful. While other repositories might already have duplicate checks, I don't know about their usage of ML to find duplicates. Based on my experience, I would guess that it isn't as accurate if it relies primarily on exact matches. CORE's duplicate check displays the "confidence" it has that a record is duplicated in the repository, which is also helpful when determining the final decision on whether a duplicate record...



Kirsten Vallee

Repository Services Manager at The University of
Chicago

[READ MORE](#)

What's next?

- Integrate with repository software.
- Automatically find duplicates and “correct” them through your data
- Automatically predict versions of papers
- Expose the version information to the scholarly graph



Conclusions

We all benefit by having a deduplicated scholarly graph

CORE can help!

Integration with repository software is central in improving the experience





Thank you!

Matteo Cancellieri
matteo.cancellieri@open.ac.uk