Leveraging Institutional Data For Author Name Disambiguation

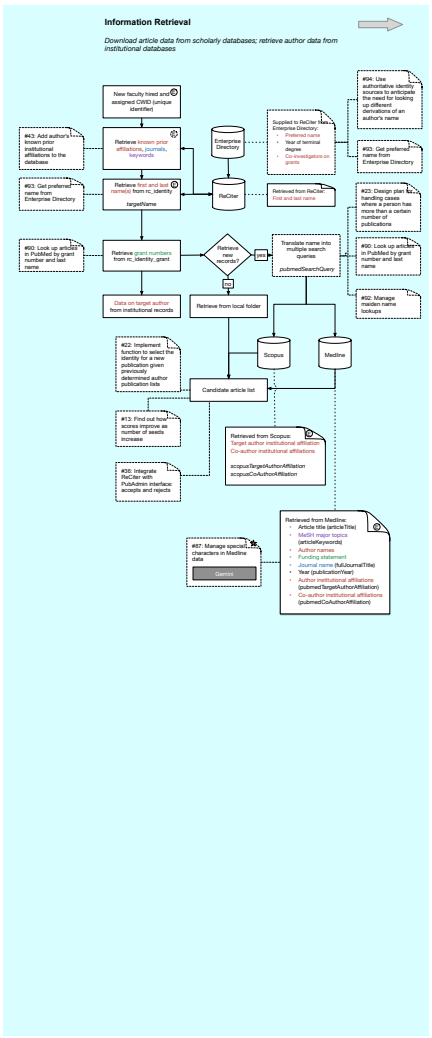
Michael Bales, Paul Albert, Jie Lin, and Stephen Johnson

The challenge: Author name disambiguation

ReCiter Architecture and Data Processing Operations

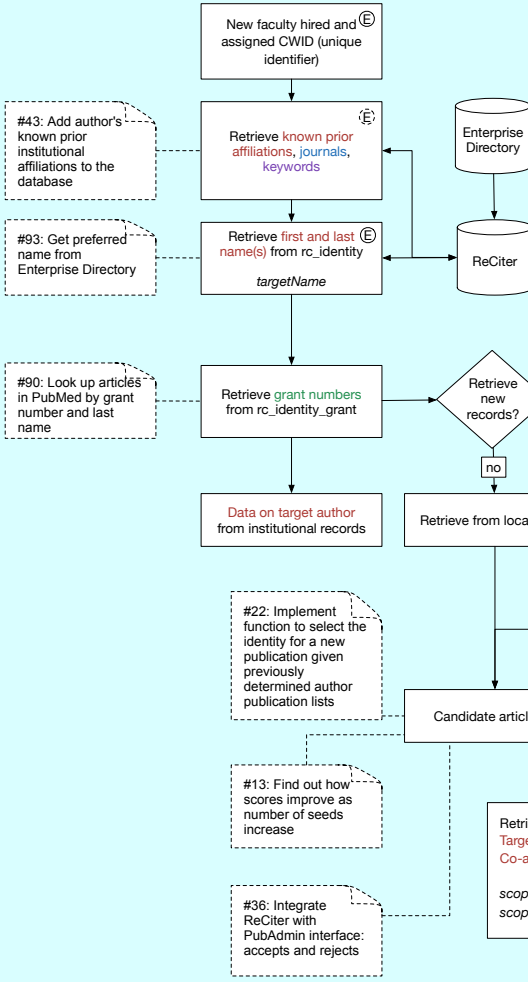
Last updated June 25th, 2015

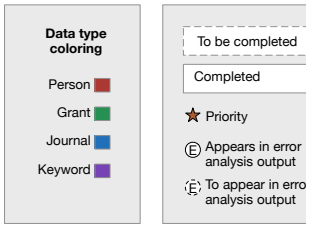
Michael Bales, Paul Albert, Jie Lin, and Stephen Johnson, Weill Cornell Medical College

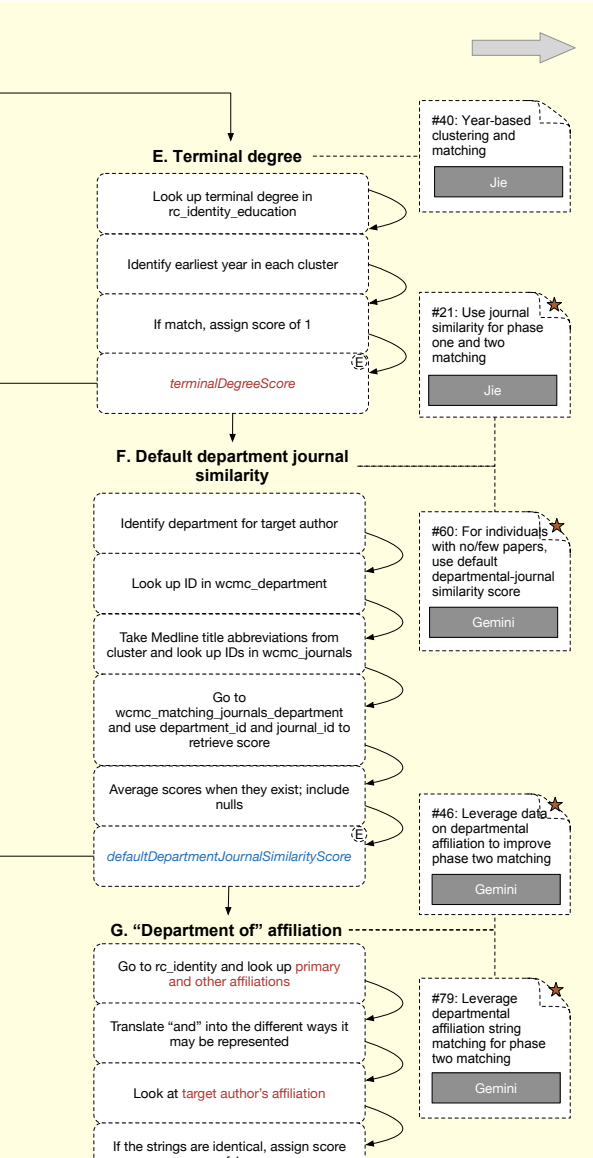


Information Retrieval

Download article data from scholarly databa
institutional databases







Output

Output preliminary calculations, scores from phase one clustering, scores from phase two matching, and clustering results for all articles input.

Information retrieval

- *status:* true/false negative/positive, according to gold standard
- *cwid:* author's institutional identifier
- *targetName:* full name as recorded in rc_identity
- *pubmedSearchQuery:* query used to identify candidate articles

Preprocessing

- *pmid:* unique ID assigned to publications in Medline
- *articleTitle:* title of article
- *fullJournalTitle:* full name as recorded in rc_identity
- *publicationYear:* year of publication
- *scopusTargetAuthorAffiliation:* affiliation of target author in Scopus
- *scopusCoAuthorAffiliation:* affiliation of co-author in Scopus
- *pubmedTargetAuthorAffiliation:* affiliation of author in PubMed
- *pubmedCoAuthorAffiliation:* affiliation of co-author in PubMed
- *articleTopicKeywords:* MeSH terms
- *targetAuthorKnownCoauthors:* last name, first initial harvested from rc_identity_grant
- *targetAuthorKnownCountry:* harvested from rc_identity_citizenship
- *targetAuthorKnownAffiliations:* institutional affiliation harvested from rc_identity_education
- *targetAuthorKnownTopicKeywords:* keywords harvested from rc_board_certification
- *targetAuthorYearTerminalDegree:* year of target author's terminal degree

Phase one clustering: clustering results and scores

- *clusterOriginator:* A "*" when an article starts a cluster; else null
- *targetAuthorNamePhaseOneScore:* Target author name in cluster versus in candidate article
- *coauthorNamePhaseOneScore:* Co-author names in cluster versus in candidate article
- *targetAuthorAffiliationPhaseOneScore:* Target author's affiliation in article versus cluster
- *coauthorAffiliationPhaseOneScore:* Co-author's affiliation in article versus cluster
- *meshMajorMatchingScore:* Overlap of MeSH major between candidate article and cluster
- *journalMatchingPhaseOneScore:* Overlap of journal titles between candidate article and cluster
- *topicKeywordMatchingPhaseOneScore:* Similarity of topic keywords between cluster and target article

#71: Improve error analysis output
Jie

#95: Output a human-readable explanation for why a publication is matched to an individual

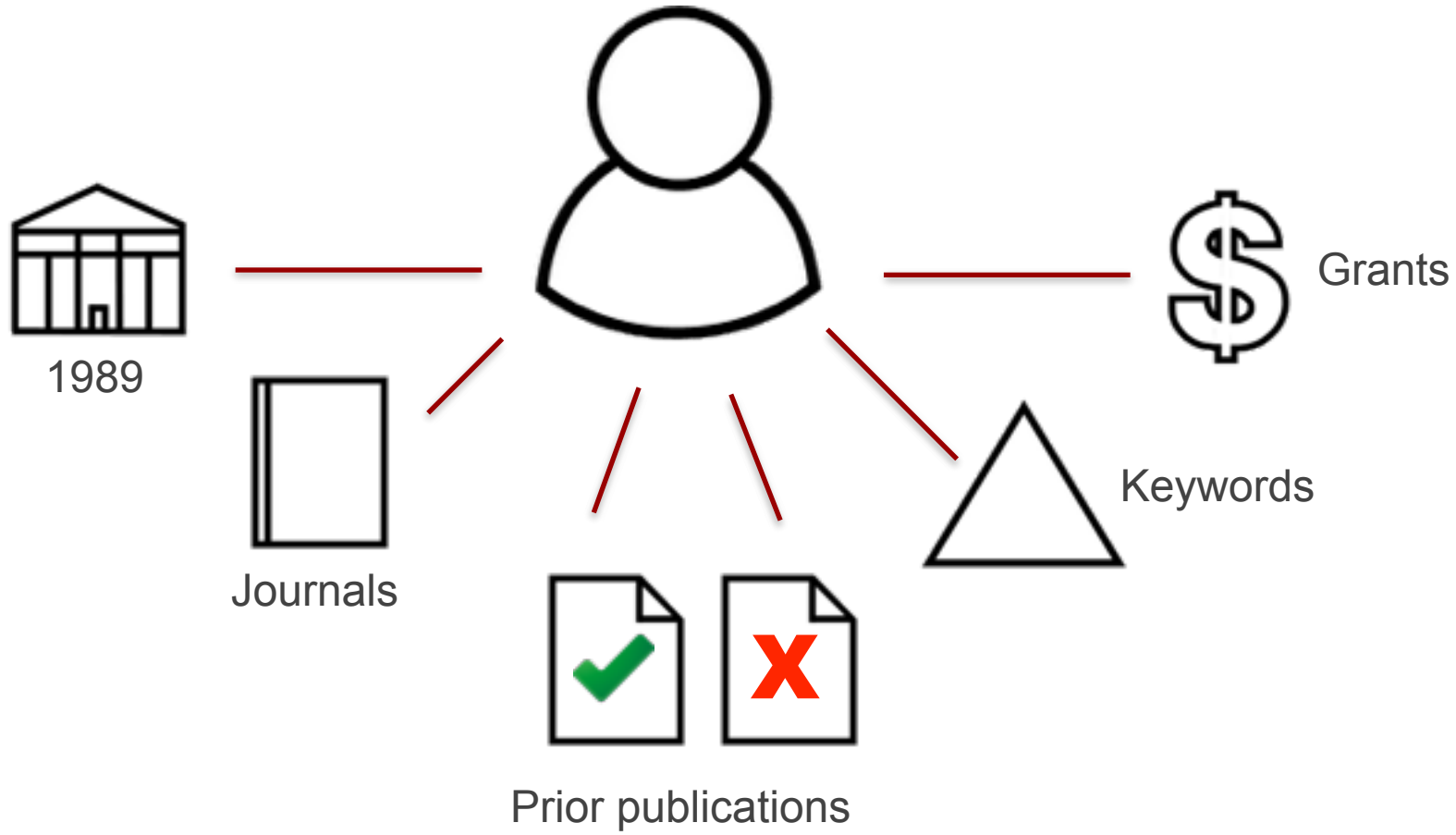
ReCiter Phases

- Information retrieval
- Preprocessing
- Phase one clustering
- Phase two matching
- Output

Use cases

1. Identify publications authored by new member of WCMC community
2. Assign newly identified publications

Evidence



ReCiter Evidence Types - Target Author

- First name, last name, middle initial
- Board certifications and clinical expertise
- Year of terminal degree
- Institutions where degrees earned
- E-mail address
- Known prior affiliations
- Geographic location of affiliation
- Grants & co-investigators

ReCiter Evidence Types, continued

- Target author prior publications
  - Journals
  - MeSH major topics
  - Co-authors
    - Names
    - Institutional affiliations
- WCMC collaborating institutions

ReCiter - Selected Evidence for Authors

Data	Source	Example	Phase 1	Phase 2
Known publication	Publications management	12923412	Yes	Yes
Job title	WOOFA	Professor of Medicine	No	Yes
Primary department	WOOFA	Medicine	No	Yes
Appointment period	WOOFA	1978 - current	No	Yes
Board certifications	Physicians profile	Cardiovascular disease	No	Yes
Citizenship	WOOFA	United States	No	Yes
Degree	WOOFA	Doctoral, 1971	No	Yes
Alma mater	WOOFA	Yale University, 1971	No	Yes
Grant	Coeus	5 U01 HL54495-10 EWOF	No	Yes

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Automatic generation of investigator bibliographies for institutional research networking systems

Stephen B. Johnson [a,*], Michael E. Bales [b], Daniel Dine [b,c], Suzanne Bakken [b,c], Paul J. Albert [d], Chunhua Weng [b,c]

[a] Department of Public Health, Weill Cornell Medical College, New York, United States
[b] Department of Biomedical Informatics, Columbia University, New York, United States
[c] The Irving Institute for Clinical and Translational Research, Columbia University, New York, United States
[d] Samuel J. Wood Library, Weill Cornell Medical College, New York, United States

ARTICLE INFO

Article history:
Received 13 December 2013
Accepted 20 March 2014
Available online 30 March 2014

Keywords:
Authorship
Bibliography as topic
MEDLINE
Natural language processing
Pattern recognition
Automated

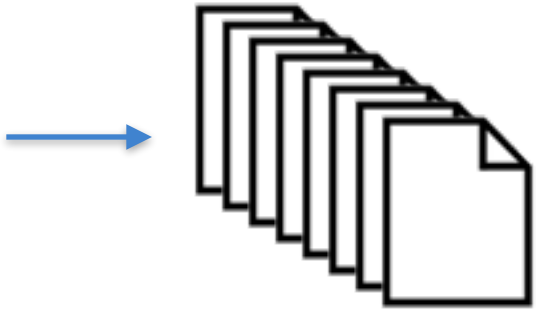
ABSTRACT

Objective: Publications are a key data source for investigator profiles and research networking systems. We developed ReCiter, an algorithm that automatically extracts bibliographies from PubMed using institutional information about the target investigators.
Methods: ReCiter executes a broad query against PubMed, groups the results into clusters that appear to constitute distinct author identities and selects the cluster that best matches the target investigator. Using information about investigators from one of our institutions, we compared ReCiter results to queries based on author name and institution and to citations extracted manually from the Scopus database. Five judges created a gold standard using citations of a random sample of 200 investigators.
Results: About half of the 10,471 potential investigators had no matching citations in PubMed, and about 45% had fewer than 70 citations. Interrater agreement (Fleiss' kappa) for the gold standard was 0.81. Scopus achieved the best recall (sensitivity) of 0.81, while name-based queries had 0.78 and ReCiter had 0.69. ReCiter attained the best precision (positive predictive value) of 0.93 while Scopus had 0.85 and name-based queries had 0.31.
Discussion: ReCiter accesses the most current citation data, uses limited computational resources and minimizes manual entry by investigators. Generation of bibliographies using named-based queries will not yield high accuracy. Proprietary databases can perform well but requite manual effort. Automated generation with higher recall is possible but requires additional knowledge about investigators.

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]



How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]

Cluster 1

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]

Cluster 1

How ReCiter Works

Ana Santos-Carvallo

↓

Santos A[AU] OR
Santos-Carvallo A[AU]

→

Cluster 1

Cluster 2

How ReCiter Works

Ana Santos-Carvallo

↓

Santos A[AU] OR
Santos-Carvallo A[AU]

→

Cluster 1

Cluster 2

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]

Cluster 1

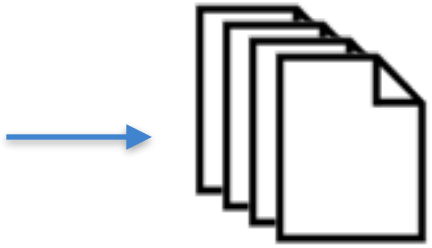
Cluster 2

Cluster 3

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]



Cluster 1

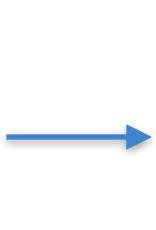
Cluster 2

Cluster 3

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]



Cluster 1

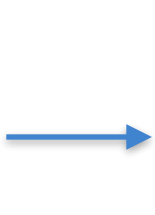
Cluster 2

Cluster 3

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]



Cluster 1

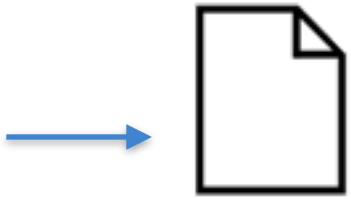
Cluster 2

Cluster 3

How ReCiter Works

Ana Santos-Carvallo

Santos A[AU] OR
Santos-Carvallo A[AU]



Weill Cornell Medical College

Cluster 1

Cluster 2

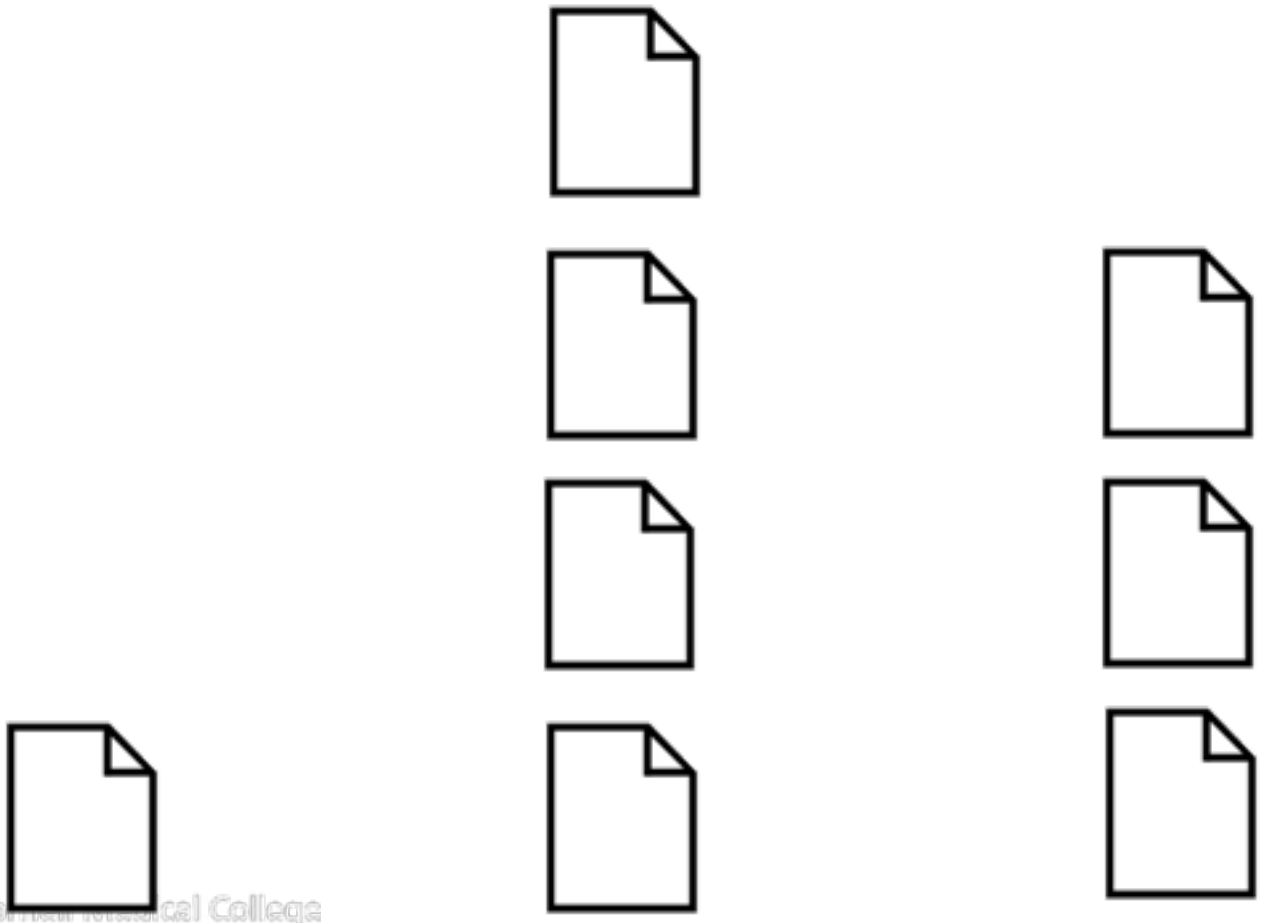
Cluster 3

How ReCiter Works

Ana Santos-Carvallo

↓

Santos A[AU] OR
Santos-Carvallo A[AU] →



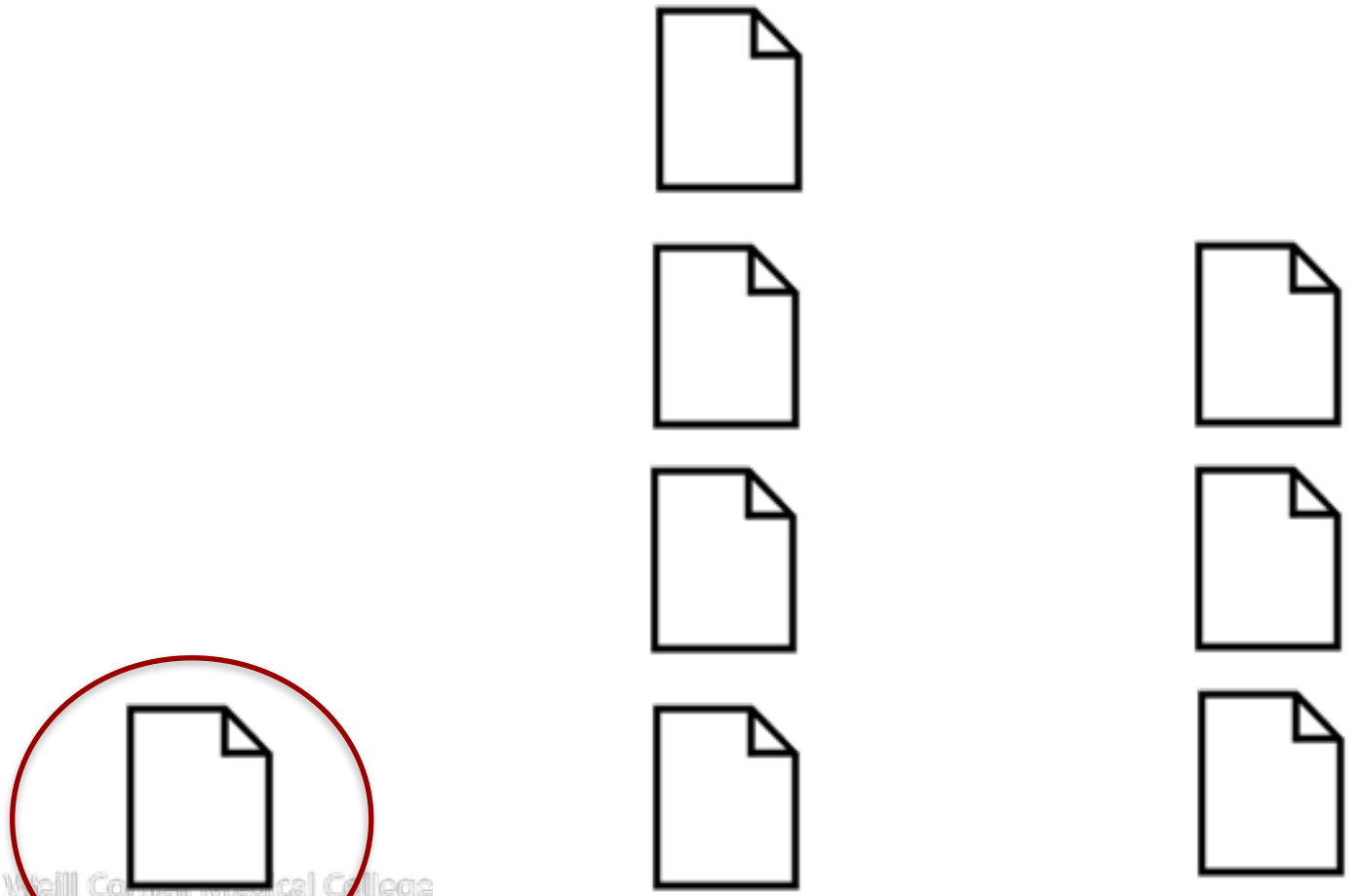
Cluster 1 Cluster 2 Cluster 3

How ReCiter Works

Ana Santos-Carvallo

↓

Santos A[AU] OR
Santos-Carvallo A[AU] →



Similarity Score: Board Certifications

Person: Jonathan W. Weinsaft

Board Certifications:
Cardiovascular Disease, Internal Medicine

Article cluster #1

- Duke Cardiovascular Magnetic Resonance Center
- Division of Cardiology, New York Methodist Hospital
- JACC. Cardiovascular Imaging

Article cluster #2

- Dept. of Ophthalmology, Mayo Clinic
- Dept. of Ophthalmology and Visual Sciences ...
- Cornea

Similarity Score: Board Certifications

Person: Jonathan W. Weinsaft

Board Certifications:
Cardiovascular Disease, Internal Medicine

Article cluster #1

- Duke Cardiovascular Magnetic Resonance Center
- Division of Cardiology, New York Methodist Hospital
- JACC. Cardiovascular Imaging

Article cluster #2

- Dept. of Ophthalmology, Mayo Clinic
- Dept. of Ophthalmology and Visual Sciences ...
- Cornea

Similarity Score: Board Certifications

Score = 0.27

Person: Jonathan W. Weinsaft

Board Certifications:
Cardiovascular Disease, Internal Medicine

Article cluster #1

- Duke Cardiovascular Magnetic Resonance Center
- Division of Cardiology, New York Methodist Hospital
- JACC. Cardiovascular Imaging

Article cluster #2

- Dept. of Ophthalmology, Mayo Clinic
- Dept. of Ophthalmology and Visual Sciences ...
- Cornea

Similarity Score: Board Certifications

Score = 0.27

Person: Jonathan W. Weinsaft

Board Certifications:
Cardiovascular Disease, Internal Medicine

Article cluster #1

- Duke Cardiovascular Magnetic Resonance Center
- Division of Cardiology, New York Methodist Hospital
- JACC. Cardiovascular Imaging

Article cluster #2

- Dept. of Ophthalmology, Mayo Clinic
- Dept. of Ophthalmology and Visual Sciences ...
- Cornea

Similarity Score: Board Certifications

Person: Jonathan W. Weinsaft

Board Certifications:
Cardiovascular Disease, Internal Medicine

Score = 0.27

Article cluster #1

- Duke Cardiovascular Magnetic Resonance Center
- Division of Cardiology, New York Methodist Hospital
- JACC. Cardiovascular Imaging

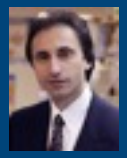
Score = 0.08

Article cluster #2

- Dept. of Ophthalmology, Mayo Clinic
- Dept. of Ophthalmology and Visual Sciences ...
- Cornea

Similarity Score: Known Co-Investigators on Grants

Person: Shahin Rafii



Grant co-investigators:
Bi-Sen Ding, Zhongwei Cao

Article cluster #1 includes:

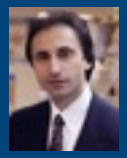
- Ding BS
- Cao Z
- Kao DI

Article cluster #2 includes:

- Kaira K
- Yasuda M
- Palefsky JM

Similarity Score: Known Co-Investigators on Grants

Person: Shahin Rafii



Grant co-investigators:
Bi-Sen Ding, Zhongwei Cao

Article cluster #1 includes:

- Ding BS
- Cao Z
- Kao DI

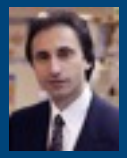
Article cluster #2 includes:

- Kaira K
- Yasuda M
- Palefsky JM

Similarity Score: Known Co-Investigators on Grants

Score = 0.45

Person: Shahin Rafii



Grant co-investigators:
Bi-Sen Ding, Zhongwei Cao

Article cluster #1 includes:

- Ding BS
- Cao Z
- Kao DI

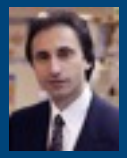
Article cluster #2 includes:

- Kaira K
- Yasuda M
- Palefsky JM

Similarity Score: Known Co-Investigators on Grants

Score = 0.45

Person: Shahin Rafii



Grant co-investigators:
Bi-Sen Ding, Zhongwei Cao

Article cluster #1 includes:

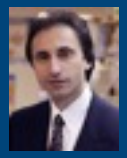
- Ding BS
- Cao Z
- Kao DI

Article cluster #2 includes:

- Kaira K
- Yasuda M
- Palefsky JM

Similarity Score: Known Co-Investigators on Grants

Person: Shahin Rafii



Grant co-investigators:
Bi-Sen Ding, Zhongwei Cao

Score = 0.45

Article cluster #1 includes:

- Ding BS
- Cao Z
- Kao DI

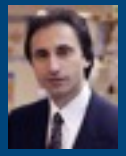
Score = 0.12

Article cluster #2 includes:

- Kaira K
- Yasuda M
- Palefsky JM

Similarity Score: Year of Terminal Degree

Person: Shahin Rafii



Year of terminal degree: 1986

Article cluster #1

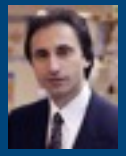
- 1972
- 1977
- 1978

Article cluster #2

- 1988
- 1990
- 1997

Similarity Score: Year of Terminal Degree

Person: Shahin Rafii



Year of terminal degree: 1986

Article cluster #1

- 1972
- 1977
- 1978

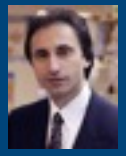
Article cluster #2

- 1988
- 1990
- 1997

Similarity Score: Year of Terminal Degree

Score = 0.04

Person: Shahin Rafii



Year of terminal degree: 1986

Article cluster #1

- 1972
- 1977
- 1978

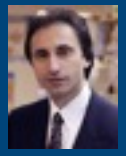
Article cluster #2

- 1988
- 1990
- 1997

Similarity Score: Year of Terminal Degree

Score = 0.04

Person: Shahin Rafii



Year of terminal degree: 1986

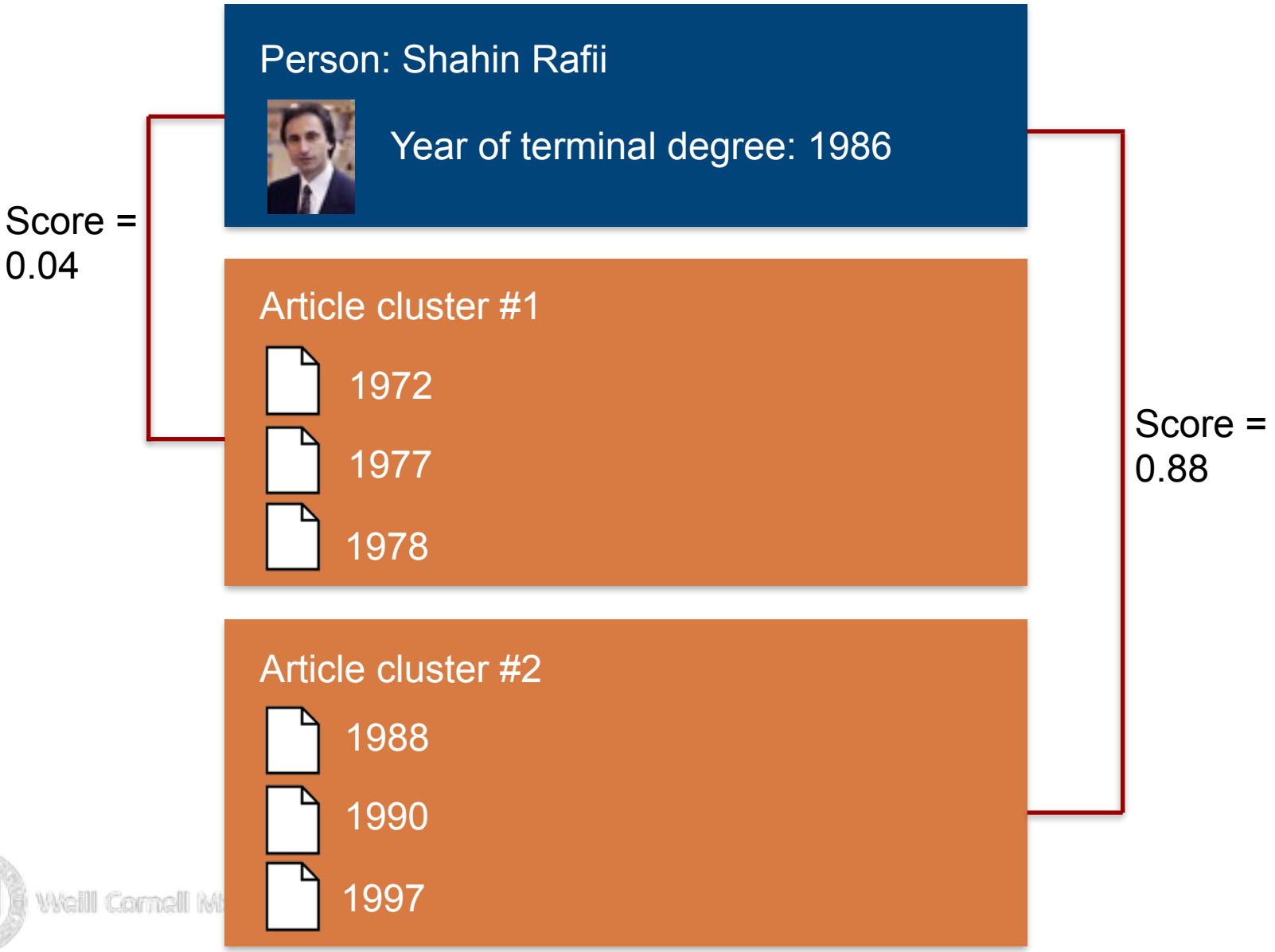
Article cluster #1

- 1972
- 1977
- 1978

Article cluster #2

- 1988
- 1990
- 1997

Similarity Score: Year of Terminal Degree



ReCiter — Edit

287 commits | 3 branches | 0 releases | 4 contributors

Branch: master ReCiter / +

issue #45 fix and implemented the JUnit test case

hanumanthaatgemini authored 2 days ago latest commit f3647931f1

src	issue #45 fix and implemented the JUnit test case	2 days ago
.classpath	Added .classpath to git	2 months ago
.gitignore	Added test for empty afid in Scopus XML.	2 months ago
.project	Added .project.	3 months ago
README.md	Formatting change	20 days ago
data.7z	Adding data.7z	2 months ago
pom.xml	Updating CSV writer.	23 days ago

README.md

ReCiter

ReCiter wiki

The wiki includes descriptions of files used for computation, an overview of error analysis, a log of performance, and use cases, among other informational material on the project.

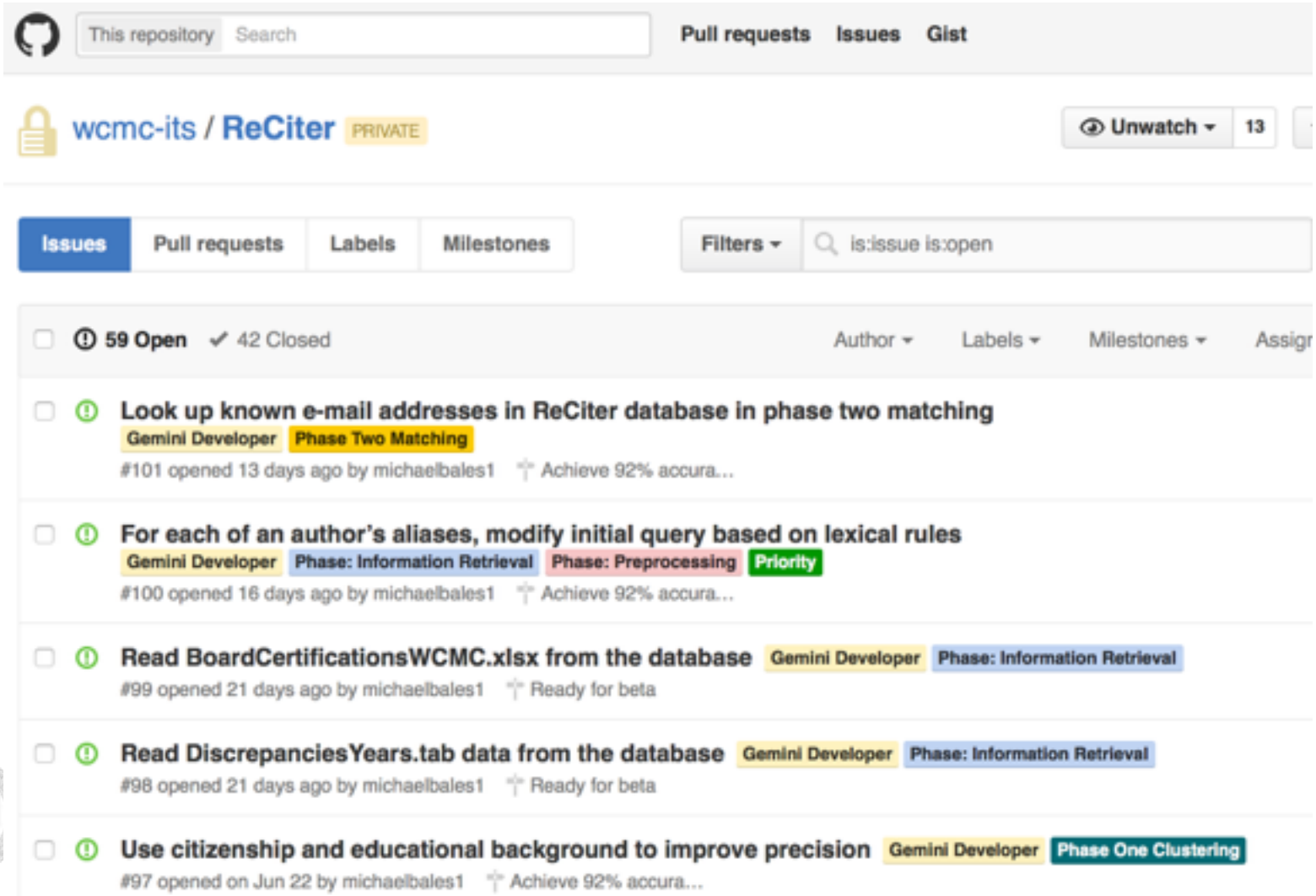
HTTPS clone URL

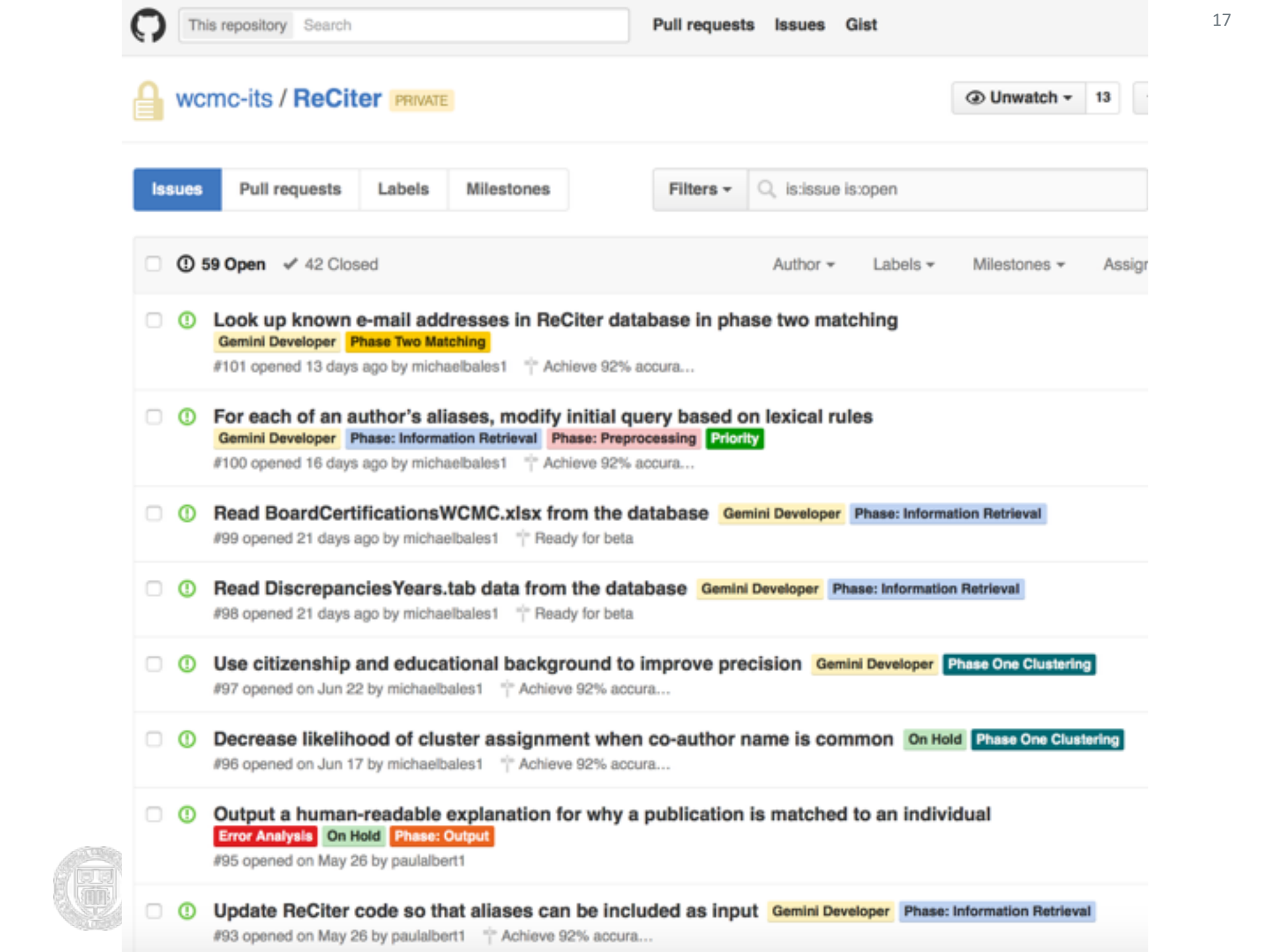
https://github.com/

You can clone with HTTPS, SSH, or Subversion.

Clone in Desktop

Download ZIP


This repository
Search
Pull requests
Issues
Gist
wcmc-its / ReCiter
PRIVATE
Unwatch
13
Issues
Pull requests
Labels
Milestones
Filters
is:issue is:open
59 Open
42 Closed
Author
Labels
Milestones
Look up known e-mail addresses in ReCiter database in phase two matching
Gemini Developer
Phase Two Matching
#101 opened 13 days ago by michaelbales1
Achieve 92% accura...
For each of an author's aliases, modify initial query based on lexical rules
Gemini Developer
Phase: Information Retrieval
Phase: Preprocessing
Priority
#100 opened 16 days ago by michaelbales1
Achieve 92% accura...
Read BoardCertificationsWCMC.xlsx from the database
Gemini Developer
Phase: Information Retrieval
#99 opened 21 days ago by michaelbales1
Ready for beta
Read DiscrepanciesYears.tab data from the database
Gemini Developer
Phase: Information Retrieval
#98 opened 21 days ago by michaelbales1
Ready for beta
Use citizenship and educational background to improve precision
Gemini Developer
Phase One Clustering
#97 opened on Jun 22 by michaelbales1
Achieve 92% accura...

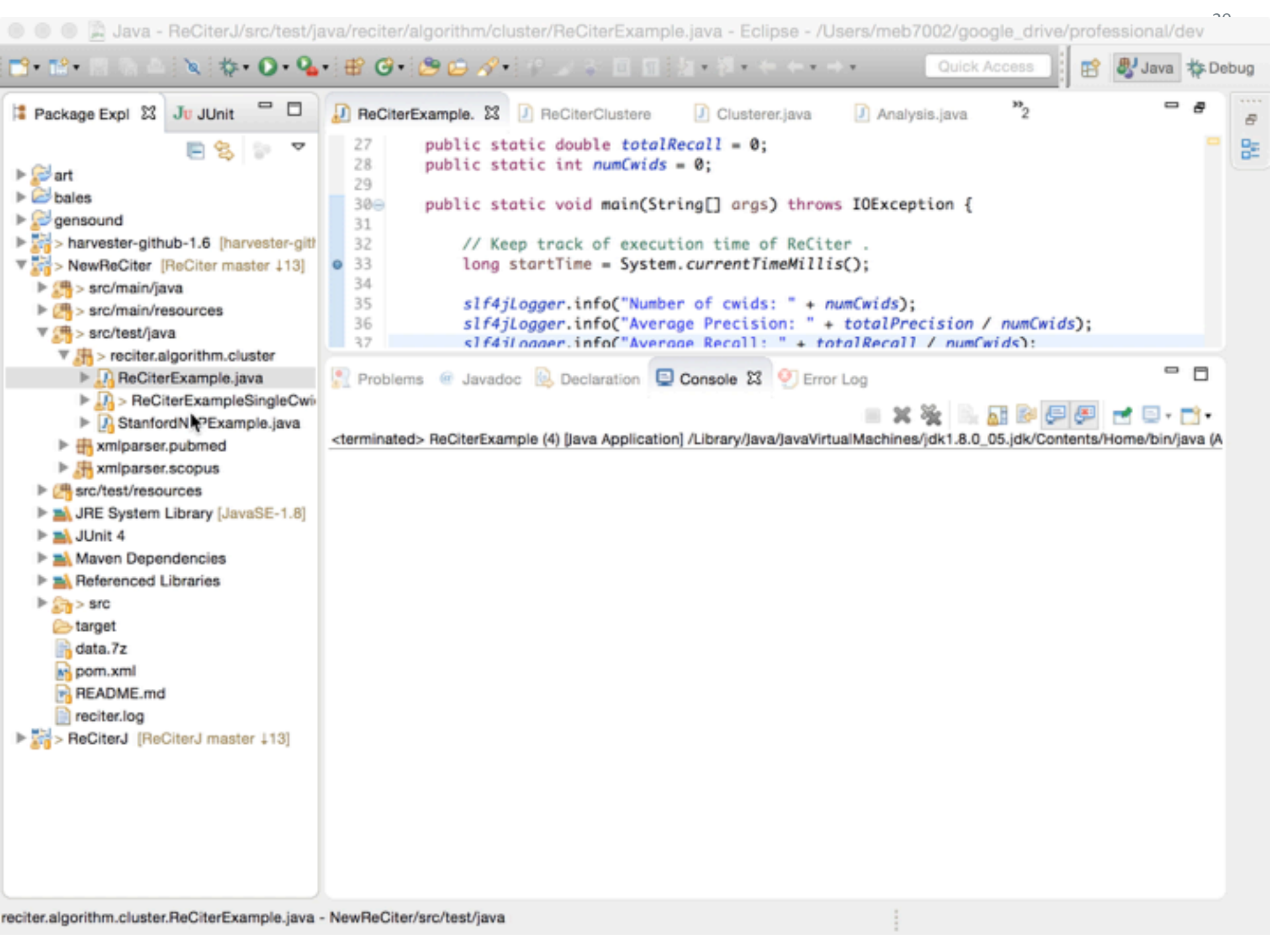

This repository
Search
Pull requests
Issues
Gist
wcmc-its / ReCiter
PRIVATE
Unwatch
13
Issues
Pull requests
Labels
Milestones
Filters
is:issue is:open
59 Open
42 Closed
Author
Labels
Milestones
Look up known e-mail addresses in ReCiter database in phase two matching
Gemini Developer
Phase Two Matching
#101 opened 13 days ago by michaelbales1
Achieve 92% accura...
For each of an author's aliases, modify initial query based on lexical rules
Gemini Developer
Phase: Information Retrieval
Phase: Preprocessing
Priority
#100 opened 16 days ago by michaelbales1
Achieve 92% accura...
Read BoardCertificationsWCMC.xlsx from the database
Gemini Developer
Phase: Information Retrieval
#99 opened 21 days ago by michaelbales1
Ready for beta
Read DiscrepanciesYears.tab data from the database
Gemini Developer
Phase: Information Retrieval
#98 opened 21 days ago by michaelbales1
Ready for beta
Use citizenship and educational background to improve precision
Gemini Developer
Phase One Clustering
#97 opened on Jun 22 by michaelbales1
Achieve 92% accura...
Decrease likelihood of cluster assignment when co-author name is common
On Hold
Phase One Clustering
#96 opened on Jun 17 by michaelbales1
Achieve 92% accura...
Output a human-readable explanation for why a publication is matched to an individual
Error Analysis
On Hold
Phase: Output
#95 opened on May 26 by paulalbert1
Update ReCiter code so that aliases can be included as input
Gemini Developer
Phase: Information Retrieval
#93 opened on May 26 by paulalbert1
Achieve 92% accura...

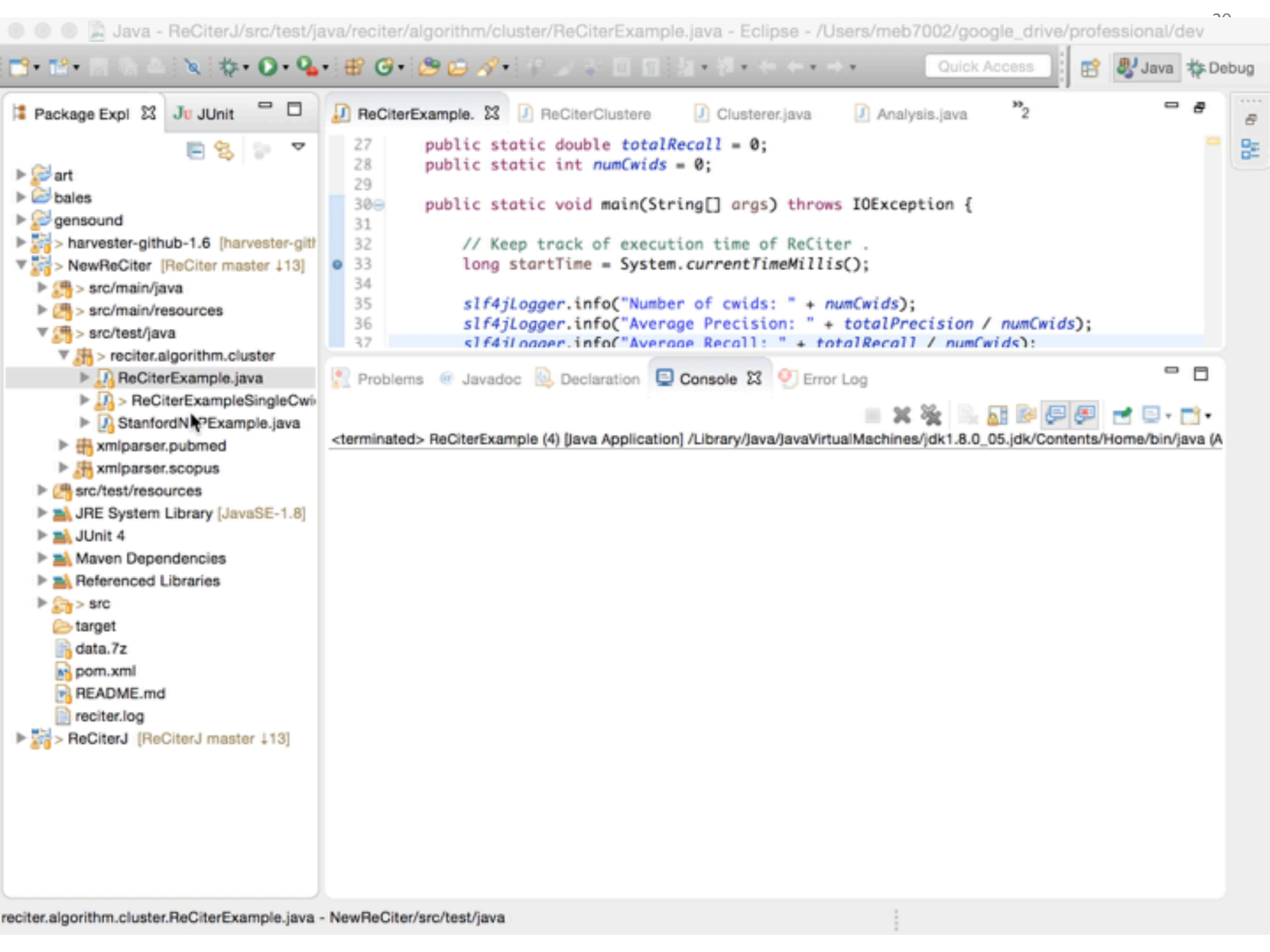
Scope

Type	Count	Priority
Active faculty	5,500	High
Active students	1,000	High
Postdocs and fellows	400	Medium
Research and staff associates	800	Medium
Alumni	5,000	Medium
Non-WCMC faculty in Graduate School	< 100	Medium
Members of the CTSC including those from CU, MSKCC, NYP, Hunter	> 10,000	Low
Inactive/historical academics	> 10,000	Low

Next steps

- Machine learning
- Open source
- You can help


Java - ReCiterJ/src/test/java/reciter/algorithm/cluster/ReCiterExample.java - Eclipse - /Users/meb7002/google_drive/professional/dev
Quick Access
Java
Debug
Package Expl
JUnit
art
bales
gensound
> harvester-github-1.6 [harvester-gith
> NewReCiter [ReCiter master ↓13]
> src/main/java
> src/main/resources
> src/test/java
> reciter.algorithm.cluster
ReCiterExample.java
> ReCiterExampleSingleCwi
StanfordNLPExample.java
xmlparser.pubmed
xmlparser.scopus
src/test/resources
JRE System Library [JavaSE-1.8]
JUnit 4
Maven Dependencies
Referenced Libraries
> src
target
data.7z
pom.xml
README.md
reciter.log
> ReCiterJ [ReCiterJ master ↓13]
ReCiterExample.
ReCiterClustere
Clusterer.java
Analysis.java
»2
27 public static double totalRecall = 0;
28 public static int numCwids = 0;
29
30 public static void main(String[] args) throws IOException {
31
32 // Keep track of execution time of ReCiter .
33 long startTime = System.currentTimeMillis();
34
35 slf4jLogger.info("Number of cwids: " + numCwids);
36 slf4jLogger.info("Average Precision: " + totalPrecision / numCwids);
Problems
Javadoc
Declaration
Console
Error Log
<terminated> ReCiterExample (4) [Java Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_05.jdk/Contents/Home/bin/java (A
reciter.algorithm.cluster.ReCiterExample.java - NewReCiter/src/test/java


Java - ReCiterJ/src/test/java/reciter/algorithm/cluster/ReCiterExample.java - Eclipse - /Users/meb7002/google_drive/professional/dev
Quick Access
Java
Debug
Package Expl
JUnit
art
bales
gensound
> harvester-github-1.6 [harvester-gith
> NewReCiter [ReCiter master ↓13]
> src/main/java
> src/main/resources
> src/test/java
> reciter.algorithm.cluster
ReCiterExample.java
> ReCiterExampleSingleCwi
StanfordNLPExample.java
xmlparser.pubmed
xmlparser.scopus
src/test/resources
JRE System Library [JavaSE-1.8]
JUnit 4
Maven Dependencies
Referenced Libraries
> src
target
data.7z
pom.xml
README.md
reciter.log
> ReCiterJ [ReCiterJ master ↓13]
ReCiterExample.
ReCiterClustere
Clusterer.java
Analysis.java
»2
27 public static double totalRecall = 0;
28 public static int numCwids = 0;
29
30 public static void main(String[] args) throws IOException {
31
32 // Keep track of execution time of ReCiter .
33 long startTime = System.currentTimeMillis();
34
35 slf4jLogger.info("Number of cwids: " + numCwids);
36 slf4jLogger.info("Average Precision: " + totalPrecision / numCwids);
Problems
Javadoc
Declaration
Console
Error Log
<terminated> ReCiterExample (4) [Java Application] /Library/Java/JavaVirtualMachines/jdk1.8.0_05.jdk/Contents/Home/bin/java (A
reciter.algorithm.cluster.ReCiterExample.java - NewReCiter/src/test/java

ReCiter Output (selected fields)

Article ID	Target author	Cluster	Articles in cluster	Cluster ultimately selected	Reference standard status
25313356	aas2004	1	4	No	True Negative
24605052	aas2004	1	4	No	True Negative
19389401	aas2004	1	4	No	True Negative
24767105	aas2004	1	4	No	True Negative
23332979	aas2004	2	2	Yes	False Positive
20489570	aas2004	2	2	Yes	True Positive

ReCiter Team

Name	E-mail
Paul Albert	paa2013@med.cornell.edu
Michael Bales	meb7002@med.cornell.edu
Jie Lin	jie265@gmail.com
Balu Mudhavathu	bam3002@med.cornell.edu
Hanumantha Rao	hat3001@med.cornell.edu

HOW MAJOR SYSTEMS TRACKING PUBLICATIONS ARE CONNECTED