ReCiter Architecture and Data Processing Operations

Last updated June 25th, 2015

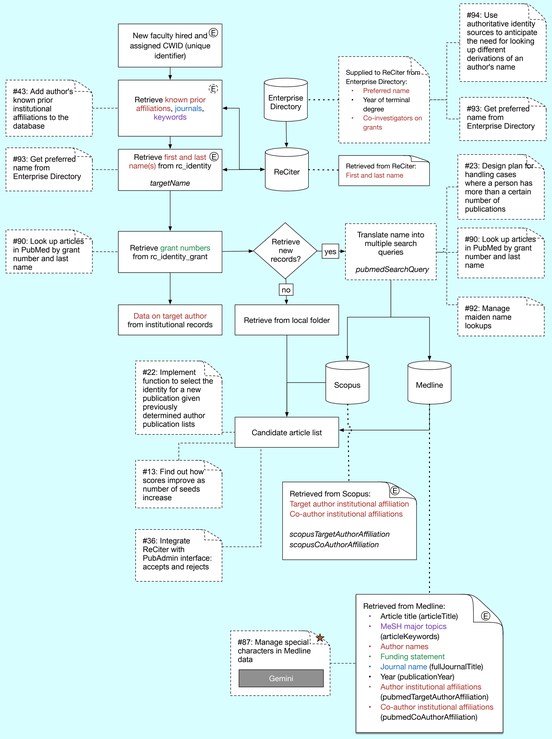
Michael Bales, Paul Albert, Jie Lin, and Stephen Johnson, Weill Cornell Medical College

Data type coloring
- Person
- Grant
- Journal
- Keyword

- To be completed
- Completed
- ★ Priority
- Ⓔ Appears in error analysis output
- Ⓔ: To appear in error analysis output

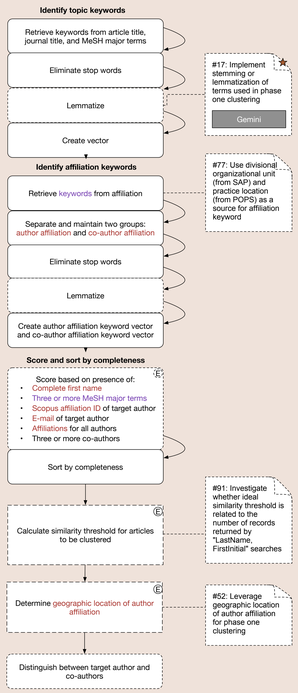
Information Retrieval

Download article data from scholarly databases; retrieve author data from institutional databases



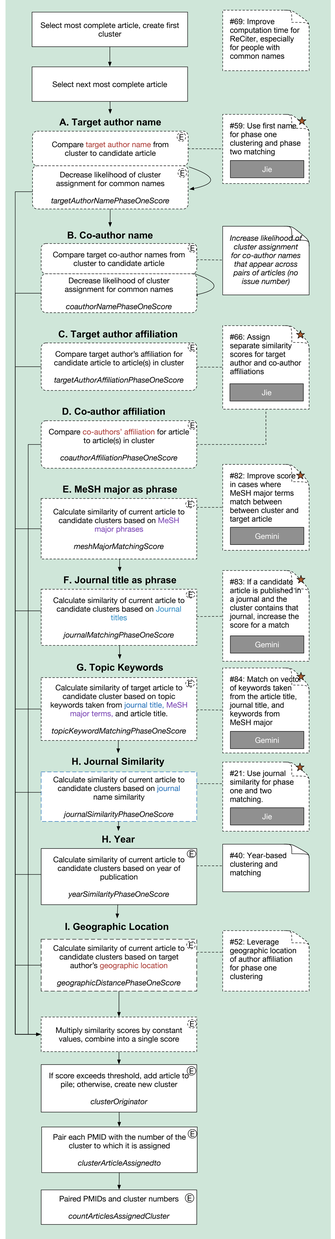
Preprocessing

Identify keywords, sort articles by completeness, calculate similarity thresholds for articles, and determine geographic location of author affiliation



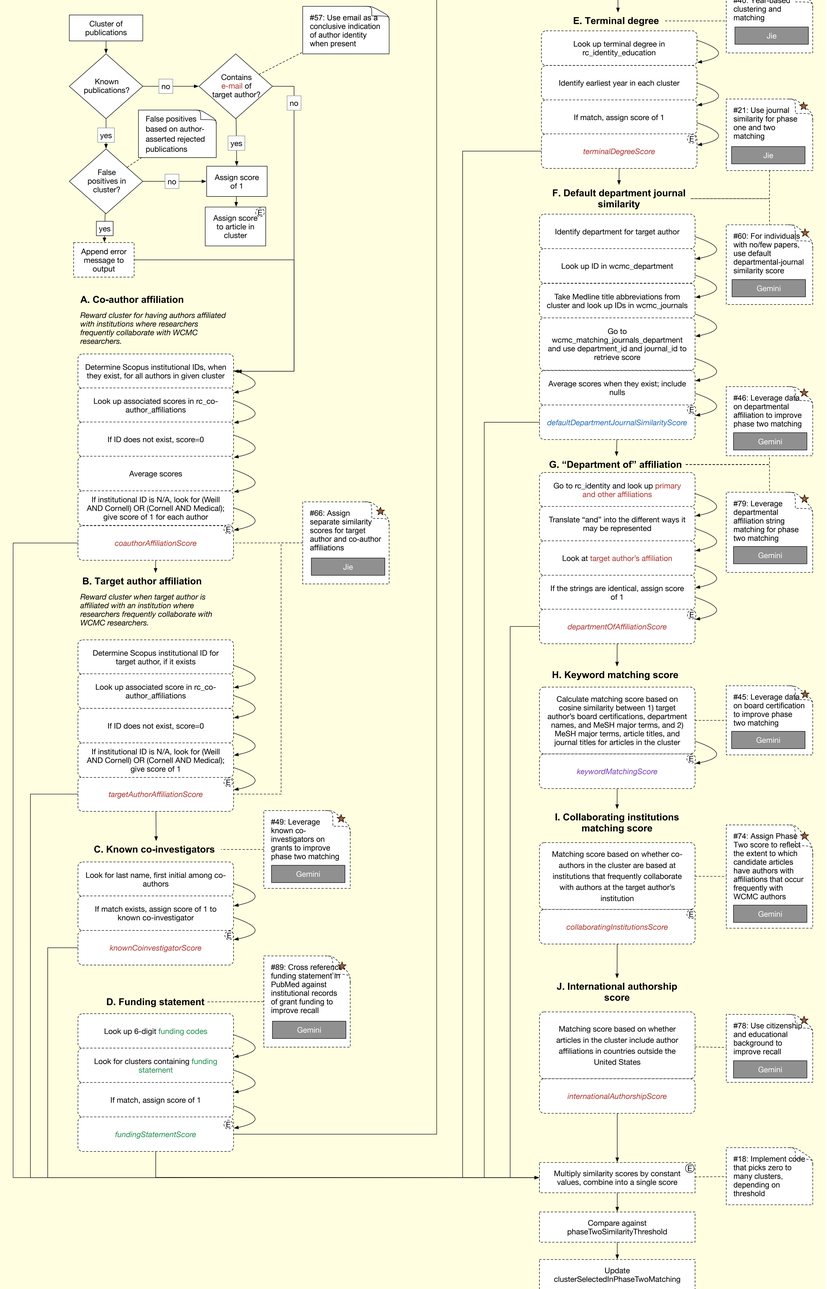
Phase 1 clustering

Sort articles into piles where each pile contains articles written by a unique author. When determining whether to add an article to an existing pile, ReCiter currently calculates cosine similarity of co-author names and keywords; the article is added to candidate pile if cosine similarity exceeds a threshold. Additional scores are to be added.



Phase 2 matching

Calculate scores for each pile; use final score to determine whether the pile corresponds to the target author.



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

#71: Improve error analysis output
Jie
#95: Output a human-readable explanation for why a publication is matched to an individual

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
journalSimilarityPhaseOneScore: Journal similarity score between target author and article cluster
yearSimilarityPhaseOneScore: Similarity of current article to candidate clusters based on year
geographicDistancePhaseOneScore: Similarity of current article to candidate clusters based on geographic distance
clusterArticleAssignedIo: Number of the cluster to which the article was assigned
countArticlesInAssignedCluster: Number of articles in the cluster to which the article was assigned

Phase two matching: matching scores

coauthorAffiliationScore: Scores are assigned to co-authors indicating whether or not they are affiliated with institutions where researchers frequently collaborate with WCMC researchers; scores are weighted by institution and averaged across co-authors (issue #47)
targetAuthorAffiliationScore: Score indicating whether target author is affiliated with institution where researchers frequently collaborate with WCMC researchers; score is weighted by institution (issue #49)
knownCoinvestigatorScore: Score indicating whether co-authors in the cluster match the names of the target author's co-investigators on grants
fundingStatementScore: Score indicating whether any articles in the candidate cluster match one or more of target author's known grant numbers (issue #89)
terminalDegreeScore Score to indicate probability the articles in the cluster belong to the target author, based on year of publication and year of target author's terminal degree (issue #40)
defaultDepartmentJournalSimilarityScore: Score to indicate probability the articles in the cluster belong to the target author, based on journal similarity (issue #21, issue #60)
departmentOfAffiliationScore: Score indicating string match between target author's departmental affiliation and affiliation of target author in Medline
keywordMatchingScore: Matching score based on cosine similarity between 1) target author's board certifications, department names, and MeSH major terms, and 2) MeSH major terms, article titles, and journal titles for articles in the cluster
collaboratingInstitutionsScore: Matching score based on whether co-authors in the cluster are based at institutions that frequently collaborate with authors at the target author's institution
internationalAuthorshipScore: Matching score based on whether articles in the cluster include author affiliations in countries outside the United States

Phase two matching: scoring results

phaseTwoSimilarityThreshold: Similarity threshold in phase two
clusterSelectedInPhaseTwoMatching: Whether the cluster was selected in phase two matching