Research on the Discovery of CORE Paper Relationships Based on Semantic Similarity and Its Semantic Services (Postprint)

Authors: Bai Linlin, Wan Ni

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose/Significance] Through a detailed dissection of the CORE paper relationship discovery process and its services, this study seeks to provide robust references for open access repositories in China concerning paper content recommendation and semantic linking. [Method/Process] The analysis examines two dimensions: the discovery process of paper relational associations based on semantic similarity, and semantic services grounded in paper relationships. Specifically, the semantic similarity-based discovery process encompasses metadata and full-text harvesting, as well as the computation of semantic similarity between papers; the semantic services based on discovered paper relationships comprise paper recommendation services and linked open data services. Finally, recommendations for CORE's application to institutional repositories in China are synthesized. [Results/Conclusion] The research reveals that the CORE system automatically harvests metadata from open access repositories via the existing OAI-PMH protocol, subsequently extracts URI fields from the metadata, and downloads full texts through the HTTP protocol. By furnishing paper recommendation services and paper linked data services based on discovered semantic relationships among papers, third-party systems can leverage the CORE dataset. These features collectively offer robust references for open access repositories in China (such as institutional repositories and open-access journals) in the realms of paper relationship recommendation and semantic linking.

Full Text

Research on CORE Paper Association Discovery and Semantic Services Based on Semantic Similarity

Bai Linlin, Wan Ni
Beijing Information Science and Technology University Library, Beijing 100192

Abstract: [Purpose/significance] This paper dissects the process and services of article association discovery in Connecting Repositories, and hopes to provide powerful reference for the recommendation and semantic linking of the content of articles in Chinese open access repositories. [Method/process] This paper analyzed the discovery process of article association based on semantic similarity and the semantic services based on article association. The discovery process of article association based on semantic similarity included metadata and full-text content harvesting, and semantic similarity calculation of article association. The semantic service based on the discovery process of article association included the CORE recommendation service and the linked open data service. And this paper summarized the application suggestions of CORE to Chinese institutional repositories. [Result/conclusion] This paper finds CORE system automatically harvests the metadata of the open access repositories through the existing OAI-PMH protocol, and further extracts the URI fields from the metadata to download the full-text through the HTTP protocol. Furtherly, providing article recommendation services and services of data linked articles based on the discovery of article semantic association enables third-party systems to utilize CORE datasets, it provides a powerful reference in recommendation and semantic linking of article association for open access repositories (such as institutional repositories and open access journals) in China.

Keywords: Connecting Repositories; semantic similarity; article association; recommendation system; linked data

---

The open access (OA) movement has promoted and facilitated global free access to scientific research outputs and the development of open access knowledge infrastructure to support functions such as content search, discovery, mining, and analysis. Currently, most open access technical infrastructures (such as institutional repositories, subject repositories, and research data repositories) are primarily based on metadata access. However, to achieve content mining and analysis functions, a crucial transformation from OA metadata integration to content integration must be realized. To this end, the European Commission-funded project "Open Access Infrastructure for Research in Europe (OpenAIRE)" established a pan-European research information platform to harvest and monitor open access research outputs funded by the European Commission and other national funders, providing rich metadata services and scientific output linking services. The project started on December 1, 2009, and has evolved through five generations (OpenAIRE, OpenAIREplus, OpenAIRE2020, OpenAIRE-Advance, OpenAIRE-Nexus). As of March 2021, the US Shared Access Research Ecosystem (SHARE) had integrated over 65.75 million research outputs from 182 data sources. France's HAL (Hyper Articles en Ligne), operated by the French National Center for Scientific Research's Institute for Research in Computer Science and Control, primarily integrates French scientific research outputs and currently contains over 2.51 million records from 168 institutions. In China, the China Academic Library & Information Sys-

tem (CALIS) has established the Chinese University Institutional Repository Alliance, which has integrated 2.86 million metadata records from 50 member institutions, while the Hong Kong Institutional Repositories (HKIR) system has integrated 426,000 records from eight universities in Hong Kong. However, these existing open access technical infrastructures only aggregate and integrate research outputs at the metadata level, without further integrating and discovering associations between papers from full-text content.

CORE (COnnecting Repositories), built in 2011 by P. Knoth from the Knowledge Media Institute of the Open University, UK, is the first system to discover associations between papers from full-text content and provide semantic services (such as recommendation services and linked data services) through various methods. The system aims to integrate open resources distributed across different systems through close collaboration with digital libraries and institutional repositories. These resources include metadata and full-text from the Directory of Open Access Journals (DOAJ), institutional repositories, and subject repositories worldwide. Based on this integration, CORE provides a series of free access services to further promote open access to scientific research outputs, making a significant contribution to the UK's open access movement and establishing its position as a aggregator of UK open access content. Since its inception, CORE has received funding from institutions including the UK Joint Information Systems Committee (JISC) and the European Commission, and has continued to develop new platform features through subsequent projects such as DiggiCORE and ServiceCORE. The DiggiCORE (Digging Into Connected Repositories) project aims to identify research community behavior patterns, research field trends, and researcher citation behaviors by analyzing large volumes of open access scientific publications using natural language processing techniques and social network analysis methods, thereby discovering high-impact papers and developing better methods for searching and browsing digital collections while forming new approaches for evaluating research and scholarly impact. The ServiceCORE project aims to further improve and perfect CORE's technical infrastructure by developing subject classification systems and knowledge discovery systems for scientific research outputs, such as building a new web service layer on top of the CORE Linked Data repository to provide programmable access to content and metadata, constructing an enhanced related resource discovery system based on text mining, and using text classification technology (support vector machines) for automatic subject-based content classification.

Based on this, a detailed analysis of CORE's paper association discovery process and the semantic services provided on this foundation can offer powerful reference and guidance for Chinese open access repositories in terms of paper content recommendation and semantic linking.

1 CORE Overview

CORE (COnnecting Repositories) is a system that aggregates OA paper metadata and full-text from around the world. As of March 2021, the system has harvested over 210 million open access papers from 13,799 institutional and subject repositories. Unlike other open access search systems that only provide metadata, CORE also integrates full-text content, ensuring free access to and download of complete scientific research outputs. Currently, CORE provides three types of services: raw data access services, content management services, and content discovery services. To improve its retrieval rate, CORE implemented CORE-MAG mapping in 2019, mapping CORE data to the Microsoft Academic Graph (MAG).

(1) Raw Data Access Services: These include CORE API, CORE Dataset, and CORE FastSync services. The CORE API provides an entry point for accessing large amounts of data in CORE, currently with two versions: a RESTful API interface providing data in XML or JSON format, and a Linked Open Data SPARQL endpoint. CORE Dataset supports users in batch downloading CORE data for processing, analysis, and mining, including paper metadata and full-text, as well as CORE-to-MAG entity mapping data. CORE FastSync enables seamless access to gold and hybrid open access papers aggregated from major publishers' non-standard systems, with data publicly shared through the FastSync protocol.

(2) Content Management Services: These include CORE Repository Dashboard and CORE Repository Edition services. CORE Repository Dashboard is a repository dashboard tool designed specifically for repository administrators, aiming to provide management and control of aggregated content. CORE Repository Edition is a suite of tools for libraries, institutional repositories, and content managers to improve the discoverability of institutional research outputs and compliance of data access.

(3) Content Discovery Services: These include CORE Recommender and CORE Discovery. CORE Recommender, as a plugin, can be used to recommend semantically similar papers between CORE and other open access repositories. CORE Discovery is a browser plugin that supports bypassing publishers to freely access papers in CORE.

2 CORE Paper Relationship Discovery Process Based on Semantic Similarity

The CORE paper relationship discovery process based on semantic similarity includes two stages: data acquisition and paper association discovery. Data acquisition primarily involves harvesting metadata records and full-text content from available open access repositories and indexing the harvested metadata and full-text. Paper association discovery primarily involves calculating and discovering semantic relationships between harvested papers through text min-

ing techniques.

2.1 Data Acquisition

2.1.1 Metadata Harvesting Metadata harvesting sources include open access repositories (institutional repositories and subject repositories) and publisher databases.

(1) Metadata from Open Access Repositories: Metadata harvesting from open access repositories is achieved through Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) requests. A successful OAI-PMH request returns an XML document containing metadata information about papers stored in the repository. The technology used in the metadata harvesting process is OCLC OAIHarvester2, a JAVA class collection toolkit for metadata harvesting through the OAI-PMH protocol.

(2) Metadata from Publisher Databases: For metadata from publisher databases that do not support the OAI-PMH protocol, the CORE team developed the CORE Publisher Connector engine, which can seamlessly access and obtain gold and hybrid open access papers from publisher databases and synchronize them through the ResourceSync protocol FastSync. Compared with the OAI-PMH protocol, which only harvests metadata for interoperability, the FastSync protocol can share any type of resource (including metadata and actual data) and provides advanced synchronization mechanisms online. The FastSync protocol is an improved version of the ResourceSync protocol, which began at the end of 2011 as a project jointly developed by the National Information Standards Organization (NISO) and the Open Archives Initiative team, funded by the Sloan Foundation and built on the OAI-PMH strategy for synchronizing metadata. This project aims to strengthen the use of modern web technology standards. CORE is one of the first companies to deploy the ResourceSync protocol to distribute large volumes of academic literature that can scale to millions of records and enable real-time updates. Currently, it has harvested 1.8 million open access papers from four major publishers: Elsevier, Springer Nature, Frontiers, and PLoS.

2.1.2 Full-Text Content Download Open access repositories provide the URL of full-text documents as part of their metadata. Therefore, full-text content acquisition is automatically downloaded from repositories through the HTTP protocol after extracting the URI field from the harvested metadata. PDF full-text downloading from open access repositories is implemented through a set of Java classes (such as the DownloadPdf class). To address slow download speeds during the download process, CORE uses BufferedStream to first download full-text content to the server, solving the problem of automatic download cancellation when remote servers respond very slowly.

2.1.3 Metadata and Full-Text Indexing After completing metadata and full-text harvesting, CORE indexes the harvested metadata and full-

text documents using Apache Lucene. The Apache Lucene project has developed three open-source search software packages: Lucene Core, Solr, and PyLucene. Lucene Core is the core sub-project, providing Java-based indexing and search technology, spell checking, hit highlighting, and advanced analysis/tokenization functions. Solr is a search platform using Lucene HTTP and JSON/Python/Ruby APIs, supporting hit highlighting, faceted search, caching, replication, and a web management interface. PyLucene is the Python port of the Lucene Core project.

2.2 CORE Paper Relationship Discovery Based on Semantic Similarity

CORE paper association discovery is completed through the semantic relationship analyzer component. This component extracts text from downloaded papers through natural language processing techniques, then identifies their association strength by calculating semantic similarity between "paper pairs." To identify and calculate semantic similarity between papers, the CORE system represents document content through vector space, converting content into a set of word vectors and finding similar documents by identifying similar vectors. The system uses the Apache Tika (PDFBox) toolkit to extract text from PDF documents, which can identify and extract metadata and text from over 1,000 different file types (such as PPT, XLS, and PDF), and calculates similarity between papers based on cosine similarity between TF-IDF vectors.

Specifically, the CORE paper relationship discovery process can be divided into the following four steps:

(1) Tokenization: Perform lexical analysis on papers downloaded by CORE to build a word dictionary T={t1,t2,…,tM}. All papers can be represented as an N × M word matrix, where N represents the number of papers and M represents the number of words formed after lexical analysis of each article, with each paper corresponding to a vector in a row of the matrix.

(2) TF-IDF Value Calculation: TF-IDF (term frequency-inverse document frequency) refers to TF*IDF, used to evaluate the importance of a word in a document collection. TF is term frequency, referring to the number of times a word appears in a single article. IDF is inverse document frequency = log2(N/DF), where DF (document frequency) represents the number of documents containing a particular word. The main idea of TF-IDF is that the importance of a word to an article primarily depends on its frequency in the document—the higher the frequency, the greater its importance to that article. Simultaneously, it is also related to the number of articles containing that word across the entire document collection; as the number of occurrences increases, it reduces the word's importance in that article. If fewer documents contain a particular term, the IDF becomes larger, and the word's discriminative power between different document categories becomes higher.

The algorithm flow is as follows: first, tokenize documents and remove stop

words; then count the occurrences of each word in individual documents and in the document collection; finally, calculate the TF-IDF value.

- **TF calculation formula:**
  Term Frequency (TF) = Number of occurrences of a word in an article / Total number of words in the article Formula (2)
- **IDF calculation formula:**
  Inverse Document Frequency (IDF) = log2(Total number of documents / Number of documents containing the word) Formula (3)

The reason for calculating inverse document frequency is to remove frequently occurring words such as "的" (de), "我们" (we), "他" (he), etc. These words have low importance for the entire document but appear frequently, potentially affecting final calculation results. Frequently occurring words cannot serve as article keywords.

- **TF-IDF value calculation:**
  TF-IDF = Term Frequency (TF) × Inverse Document Frequency (IDF) Formula (4)

(3) Sorting: Sort the TF-IDF values of article words, select words with relatively large TF-IDF values, merge them into a set, calculate the word frequency of each article for words in this set, generate respective word frequency vectors for each article, and then calculate similarity between article word frequency vectors.

(4) Similarity Calculation: Many methods exist for calculating similarity between two vectors, such as cosine similarity, dice coefficient, or Jaccard methods, with some studies employing dimensionality reduction algorithms before similarity calculation to improve performance. CORE adopts the most standard similarity calculation method: calculating cosine similarity based on TF-IDF vectors. Compared with other similarity calculation methods, the cosine similarity method for TF-IDF vectors is relatively mature and has been used in automatic link generation systems. The complete formula is as follows:

$$\text{sim}(x,y) = \cos(x,y) = (x \cdot y) / (||x|| \times ||y||) \qquad \text{Formula (5)}$$

The similarity between vectors can be judged by the size of the angle. The smaller the angle, the larger the cosine value, and the more similar they are.

3 Semantic Services Based on Discovered CORE Paper Relationships

CORE provides users with similar paper recommendation services and linked open data services based on discovered paper relationships. The similar paper recommendation service is provided in the form of CORE Recommender plugins and CORE API. The linked open data service refers to CORE publishing paper similarity data as linked data and registering it in the Linked Data Cloud.

3.1 CORE Recommendation Service

In April 2013, CORE first released a recommendation system for Eprints repositories called CORE Widget, published in the Eprints Bazaar, a store for installing Eprints add-on components and patches. In October 2016, CORE launched a new version with many improvements and upgrades to the original "CORE Widget" recommendation system, renaming it CORE Recommender. The upgraded recommendation system not only supports recommending similar papers within CORE but can also be deployed in other repositories and journal systems to recommend similar papers. For Eprints repositories, installation is as simple as downloading from the Eprints Bazaar; for other repositories (Dspace, Fedora, OJS), installation only requires inserting a snippet of Javascript code. Currently, it has been used in multiple repositories, such as the University of Strathclyde's institutional repository Strathprints, the Latin American institutional repository network LA Referencia, the Russian State Vocational Pedagogical University institutional repository, and the preprint repository arXiv.

To improve the quality of recommended similar papers, CORE Recommender employs multiple filters and crowdsourcing mechanisms to screen recommended papers, such as providing only open access papers, including only papers with at least a minimum set of metadata attributes, and including papers with thumbnails. However, in some cases, CORE Recommender may provide irrelevant or even incorrect recommendations. To address this, CORE provides users with feedback buttons for error reporting. If users report that recommended papers are inappropriate, CORE adds these papers to a blacklist and no longer displays them in recommendation lists (see Figure 1 [Figure 1: see original paper]).

CORE Recommender has two usage modes. The first mode is deployed as a recommendation system within the CORE system itself, recommending similar papers to the currently accessed paper (see Figure 1). The second mode is installed and integrated as a recommendation plugin into repository systems or journal systems. When a user visits a paper page in a repository, the plugin sends information about the accessed entry to CORE, which returns a list of similar papers. Currently, two forms of similar lists are provided: one from similar papers in the CORE repository, and the other from similar papers in the repository being accessed by the user (see Figure 2 [Figure 2: see original paper]).

3.2 Linked Data Service

In 2011, CORE released over 3 million RDF triples generated based on similarity calculations of more than 400,000 full-text papers, achieving linked data publication of paper similarity metadata to facilitate flexible access by third parties. In the process of publishing paper similarity relationships as linked data, CORE selected the Sesame platform as the triplestore for publishing linked data. The following sections elaborate on CORE's data model and implementation mechanism for publishing paper relationships as linked data.

3.2.1 CORE Data Model Following linked data principles, when publishing data as linked data, CORE reuses existing vocabularies or ontologies as much as possible to describe data, making it easier for the external world to integrate new data with existing datasets and services. CORE uses the MuSim Similarity Ontology, the Bibliographic Ontology (BIBO), and its own constructed ontology (core) to represent relationships between papers in the CORE repository.

The MuSim Similarity Ontology was collaboratively developed by K. Jacobson from the Beautiful Digital Music Center, Y. Raimond from the BBC, T. Gängler from Dresden University of Technology, and others. Initially designed to represent similarity between music, it can also be applied to other domains to represent similarity and association between two things, facilitating recommendation and discovery of related items in different contexts. This ontology contains 5 classes and 13 properties. In CORE, it is primarily used as the subject of a document, with properties including document type (rdf:type), similar papers (MuSim:element), OAI identifiers (core:hasOAIRepositoryIdentifier, core:hasOAIIdentifier), similarity calculation methods between papers (MuSim:method), and similarity weights (MuSim:weight) (see Figure 3 [Figure 3: see original paper] and Figure 4 [Figure 4: see original paper]).

The BIBO Bibliographic Ontology was collaboratively developed by F. Giasson and B. D'Arcus for describing bibliographic references and citations in the Semantic Web, with strong extensibility allowing other vocabularies to be mixed in, such as FOAF, DC, and Event vocabularies. In CORE, BIBO classes and properties are used to semantically describe paper document types, authors, etc.

3.2.2 Sesame Linked Data Implementation Mechanism Sesame is an open-source framework for querying and analyzing RDF data, originally created by the Dutch software company Aduna and inherited by the Eclipse RDF4J project in May 2016. It primarily runs as two Java Web applications: OpenRDF Sesame Server and OpenRDF Workbench. OpenRDF Sesame Server accesses Sesame repositories through HTTP, providing only server log information viewing functions without any user-facing features. OpenRDF Workbench provides user-facing query, browse, update, and export functions through a web interface.

Since its inception, CORE has used the Tomcat Web server as its application container, which supports Java Servlets and JSP technology. Therefore, CORE deploys Sesame's two components—OpenRDF Sesame Server and OpenRDF Workbench—as Java Servlet applications on the Tomcat Web server.

Specifically, Sesame is divided into three layers:

(1) Storage and Inference Layer: Sesame's storage and inference functions are implemented through the SAIL (Storage and Inference Layer) API, which abstracts from underlying storage repositories and supports in-memory triplestores, on-disk triplestores, and relational database storage. Two separate Servlet packages manage access to these triplestores on permanent servers.

(2) Linked Data Conversion Layer: The linked data conversion process is implemented through the Sesame Rio (RDF) package. Sesame Rio is a simple API consisting of Java-based RDF parsers and writers for inputting/outputting RDF data. Users can easily extend it by placing parsers and writers on the Java classpath when running applications.

(3) Linked Data Query and Access Layer: These functional modules can be accessed through Sesame's Access API, which consists of two independent parts: Repository API and Graph API. Repository API provides high-level access to Sesame repositories, such as querying, storing RDF files, and extracting RDF. Graph API provides more fine-grained support for RDF operations, such as adding and removing individual statements and creating small RDF models directly from code. These two APIs are functionally complementary and are often used together in practice. Sesame supports two query languages: SPARQL and SeRQL, and can also add free text search functionality through LuceneSail.

4 CORE's Application Implications for China's Institutional Repositories

By integrating OA paper metadata and full-text from around the world, CORE provides recommendation services based on paper similarity and semantic services based on linked data, achieving an effective transformation from OA metadata integration to content integration, improving resource visibility and access rates, and advancing traditional OA repository integration systems. This offers novelty and reference significance for the development and improvement of China's institutional repositories, which are still in their early stages. This section summarizes CORE system's implications for China's institutional repositories from three aspects: paper relationship discovery process, paper recommendation services, and linked data services.

In terms of paper relationship discovery, CORE first harvests metadata, further extracts URI fields from the harvested metadata, and then automatically downloads full-text from repositories through the HTTP protocol. On this basis, it extracts text from downloaded papers through natural language processing techniques and calculates similarity between paper pairs to identify paper relationships. Currently, China's institutional repository integration systems have achieved metadata-level harvesting but have not implemented full-text acquisition. However, since the harvested metadata fields already contain URI fields, subsequent full-text acquisition through URIs is needed, and the obtained full-text should be processed through natural language processing techniques to extract text and calculate similarity between papers to identify paper relationships.

In terms of paper semantic recommendation services, CORE implements paper recommendations by deploying its developed CORE Recommender plugin within CORE or other repositories. China's institutional repositories can learn from this approach by developing recommendation service systems or introduc-

ing CORE Recommender plugins deployed in institutional repositories to recommend similar papers to users.

In terms of linked data services, CORE publishes paper data as linked data by utilizing existing vocabularies—the MuSim Similarity Ontology, BIBO Bibliographic Ontology, and the Sesame platform—facilitating better semantic linking for users. China can model institutional repository data, reuse existing mature vocabularies to describe data as much as possible, and utilize open-source linked data publishing tools and platforms to semantically organize and publish literature resources in institutional repositories, thereby improving resource discoverability and visibility.

5 Issues in CORE's Paper Relationship Discovery Process and Services

CORE also faces many problems and challenges in its paper relationship discovery process and related services, with specific solutions as follows:

(1) Full-text content download: This primarily involves file download speed and data storage cost issues. To address download speed, CORE uses BufferedStream to first download full-text content to the Open University server, solving the problem of automatic download cancellation when remote servers respond very slowly. Regarding data storage costs, since CORE needs to download data from many open access repositories, the system requires substantial disk space. To perform system backups and allow rapid response, it chooses Serial Attached SCSI (SAS) disks.

(2) Text extraction: CORE tested three PDF text extraction systems: iText, Apache Tika (PDFBox), and pdftotext, ultimately finding that although Apache Tika's extraction speed was very slow, the quality of extracted text was high. Finally, by using BufferedStreams for pre-buffering, CORE managed to speed up the extraction process.

(3) Similarity calculation: To discover relevant papers within a reasonable time frame, a massive paper combination problem must be addressed. CORE developed a new heuristic method that uses document frequency cutoff criteria to reduce the number of combinations to consider. Considering computational result quality issues, CORE developed its own TextAnalyzer and TextFilter on the Lucene library to filter mathematical formulas, numbers, and other types of noise data.

6 Conclusion

This paper analyzed CORE's paper metadata and full-text acquisition process, paper relationship discovery through semantic similarity calculation, and services provided based on discovered paper semantic relationships, offering powerful reference for Chinese open access repositories in terms of paper relationship discovery process, paper recommendation services, and linked data services.

However, CORE also faces issues such as slow download speeds, large storage overhead, slow PDF text extraction, and similarity calculation accuracy, which require further in-depth research.

References

[1] Openaire-history [EB/OL]. [2021-03-01]. https://www.openaire.eu/openaire-history.
[2] SHARE [EB/OL]. [2021-02-27]. https://share.osf.io/.
[3] The open archive HAL [EB/OL]. [2021-03-01]. https://hal.archives-ouvertes.fr/.
[4] 中国高校机构知识库联盟 [EB/OL]. [2021-03-01]. http://chair.calis.edu.cn/.
[5] Hong Kong Institutional Repositories (HKIR) [EB/OL]. [2021-03-01]. https://library.tu.ac.th/tu-digital-collections/hong-kong-institutional-repositories-hkir.
[6] CORE – Aggregating the world's open access research papers [EB/OL]. [2021-03-01]. https://core.ac.uk/.
[7] COnnecting Repositories [EB/OL]. [2021-03-01]. https://en.wikipedia.org/wiki/COnnecting_{REpositories
[8] Knowledge Media Institute [EB/OL]. [2021-03-01]. https://www.open.ac.uk/research/kmi/.
[9] CORE | Jisc [EB/OL]. [2021-03-01]. https://www.jisc.ac.uk/core#.
[10] Digging into Connected Repositories (DiggiCORE) [EB/OL]. [2021-03-01]. https://diggingintodata.org/awards/2011/project/digging-connected-repositories-diggicore.
[11] Data Providers [EB/OL]. [2021-03-01]. https://core.ac.uk/dataproviders.
[12] CORE Services [EB/OL]. [2021-03-01]. https://core.ac.uk/services.
[13] CORE Dataset [EB/OL]. [2021-03-01]. https://core.ac.uk/documentation/dataset/.
[14] Connecting Repositories (CORE) | Digging Into Data [EB/OL]. [2021-03-01]. https://diggingintodata.org/repositories/connecting-repositories-core.
[15] Open Archives Initiative Protocol for Metadata Harvesting [EB/OL]. [2021-03-01]. http://www.openarchives.org/pmh/.
[16] OAIHarvester2 [EB/OL]. [2021-03-01]. https://www.oclc.org/research/activities/oaiharvester2.html.
[17] Technical standards [EB/OL]. [2021-03-01]. https://blog.core.ac.uk/2011/03/.
[18] Releasing 1.8 million open access publications from publisher systems for text and data mining [EB/OL]. [2021-03-01]. https://blogs.lse.ac.uk/impactofsocialsciences/2018/03/22/releasi 1-8-million-open-access-publications-from-publisher-systems-for-text-and-data-mining/.
[19] Java 文件流 BufferedStream [EB/OL]. [2021-03-01]. https://blog.csdn.net/mariofei/article/details/5119505
[20] Apache Lucene [EB/OL]. [2021-03-01]. http://lucene.apache.org/.
[21] KNOTH P, ROBOTKA V, ZDRAHAL Z. Connecting repositories in the open access domain using text mining and semantic data [C]// International conference on theory and practice of digital libraries: research and advanced technology for digital libraries. Berlin: Springer, 2011: 483-487.
[22] Apache Tika [EB/OL]. [2021-03-01]. https://tika.apache.org/.
[23] FRANCINE C, AYMAN F, THORSTEN B. Multiple similarity measures and source-pair information in story link detection [C]// Proceedings of the human language technology conference of the North American Chapter of

the Association for Computational Linguistics: HLT-NAACL 2004. Boston: Association for Computational Linguistics, 2004: 313-320.
[24] CORE - Semantic Similarity of Open Access publications [EB/OL]. [2021-03-01]. https://lod-cloud.net/dataset/core.
[25] The EPrints Bazaar [EB/OL]. [2021-03-02]. https://bazaar.eprints.org/.
[26] CORE Recommender [EB/OL]. [2021-03-03]. https://core.ac.uk/services#recommender.
[27] Implementing the CORE Recommender in Strathprints: a "white-hat" improvement to promote user interaction [EB/OL]. [2021-03-03]. https://blog.core.ac.uk/2017/10/31/implementing-the-core-recommender-in-strathprints-a-whitehat-improvement-to-promote-user-interaction/.
[28] LA Referencia integrates CORE Recommender in its services [EB/OL]. [2021-03-03]. https://blog.core.ac.uk/2019/11/20/la-referencia-integrates-core-recommender-in-its-services/.
[29] CORE Recommender installation for DSpace [EB/OL]. [2021-03-03]. https://blog.core.ac.uk/2020/03/12/core-recommender-installation-for-dspace/.
[30] CORE Recommender now supports article discovery on arXiv [EB/OL]. [2021-03-03]. https://blog.arxiv.org/2020/10/15/core-recommender-now-supports-article-discovery-on-arxiv/.
[31] Sesame (framework) – Wikipedia [EB/OL]. [2021-03-06]. https://en.wikipedia.org/wiki/Sesame_(framewo
[32] The Similarity Ontology [EB/OL]. [2021-03-04]. http://grasstunes.net/ontology/similarity/0.2/musim.htm
[33] D'ARCUS B, GIASSON F. Bibliographic ontology specification [EB/OL]. [2021-03-05]. http://bibliontology.com/.
[34] Eclipse RDF4J – a Java framework for RDF [EB/OL]. [2021-03-10]. http://rdf4j.org/.
[35] Overview (OpenRDF Sesame 4.1.2 API) [EB/OL]. [2021-03-15]. http://archive.rdf4j.org/javadoc/sesame-4.1.2/.
[36] Apache Tomcat® [EB/OL]. [2021-03-15]. http://tomcat.apache.org/.
[37] Chapter 1. Introduction: what is Sesame? [EB/OL]. [2021-03-17]. https://poc.vl-e.nl/distribution/manual/sesame-1.2.3/ch01.html.
[38] The SAIL API [EB/OL]. [2021-03-18]. http://docs.rdf4j.org/sail/.

Author Contributions:
Bai Linlin: Responsible for data acquisition, research outline determination, and paper writing;
Wan Ni: Responsible for paper revision.

Note: Figure translations are in progress. See original paper for figures.