Postprint of Applied Research on Linked Data for Similar Literature Discovery in Academic Resource Networks

Authors: Zhao Yiping, Bi Qiang

Date: 2017-10-11T00:00:00+00:00

Abstract

【Purpose】To leverage the advantages of linked data—including machine readability, semantic representation, relational description, and web resource attributes—to compensate for deficiencies in information organization within academic resource networks, thereby supporting the discovery of similar literature.

【Method】 The Latent Semantic Analysis (LSA) method is employed to calculate the overall similarity of documents published on academic resource networks. Hierarchical clustering is utilized to determine similarity thresholds for filtering, thereby generating a document relationship matrix. Based on this, dynamic document technology is used to construct linked data for academic resource networks to support semantic retrieval of related literature.

【Results】 Linked data for academic resource networks with a similar literature query function has been preliminarily implemented, enabling convenient retrieval of literature highly relevant to any given document and facilitating efficient discovery of similar literature.

【Limitations】The discovery of similar literature in academic resource networks is currently realized solely from a statistical perspective; further research is needed to achieve deeper discovery by leveraging the knowledge structure, semantic connotations, and organizational methods of document collections.

【Conclusion】 The Latent Semantic Analysis method for calculating document similarity can effectively identify similar documents. Recording associations between similar literature in linked data supports semantic retrieval to obtain precise similar literature and can substantially reduce the latency of real-time similarity computation.

Full Text

Using Linked Data to Discover Similar Documents in Academic Resource Websites

Zhao Yiping, Bi Qiang
(School of Management, Jilin University, Changchun 130022, China)

Abstract

[**Objective**] This study leverages the machine-readability, semantic representation, relational description, and web resource attributes of linked data to address information organization deficiencies in academic resource websites and support the discovery of similar documents. [**Methods**] We employ latent semantic analysis to calculate the overall similarity of documents published on academic resource websites, determine similarity thresholds through hierarchical clustering for filtering, generate a document relation matrix, and subsequently utilize dynamic document technology to construct linked data for academic resource networks to support semantic retrieval of related literature. [**Results**] We preliminarily implemented linked data for academic resource websites with similar document query functionality, enabling convenient access to highly relevant literature for any given document and facilitating efficient discovery of similar papers. [**Limitations**] Our approach discovers similar documents in academic resource networks solely from a statistical perspective; deeper similarity detection utilizing document collection knowledge systems, semantic connotations, and organizational methods requires further investigation. [**Conclusions**] Latent semantic analysis effectively identifies similar documents through similarity computation. Recording similar document relationships in linked data supports semantic retrieval to obtain precise results while substantially reducing real-time similarity calculation latency.

Keywords: Linked Data; Latent Semantic Analysis; Academic Resource Websites; Similarity
Classification Number: G354

Academic resource websites are online spaces where scholars in specific fields publish academic insights and achievements and exchange ideas, containing extremely rich disciplinary information resources. Examples include ScienceNet (www.sciencenet.cn), COS (cos.name), Muchong (emuch.net), and the Chinese W3C Alliance (w3china.org). These disciplinary information resources emerge from academic communication, exhibiting strong timeliness, conciseness, and non-systematic characteristics that require systematic organization to maximize their academic facilitation functions. Digital resource aggregation constructs interconnected, multi-dimensional, and multi-layered resource systems by strengthening semantics and discovering relationships, forming knowledge networks that integrate conceptual themes, disciplinary content, and research objects. This paper adopts the knowledge reorganization concept of clustering

from digital resource aggregation, conducting semantic mining of internal document characteristics in academic resource websites when target digital resource categories, architecture, and specialized vocabularies are unknown, and revealing similar documents through relational analysis and linked data construction approaches.

Related Research Progress

Similar document discovery primarily relies on measuring inter-document similarity, with research directions mainly including: (1) document similarity calculation methods based on statistical approaches such as co-word analysis and vector space models; and (2) semantic similarity calculation methods based on knowledge system semantic understanding.

Statistical-based document similarity measurement methods primarily calculate the frequency of main vocabulary from a document appearing in other documents, offering low-cost, high-efficiency advantages. Notably, when document collection capacity requirements are modest but measurement precision is high, these methods can complete calculations without domain vocabulary assistance. Magerman et al. used text mining technology combining latent semantic analysis and vector space models to evaluate similarity between patents and scientific publications, manually verifying document similarity relationships to ensure they weren' t affected by high-frequency terms, applying term frequency-inverse document frequency weighting corrections to obtain accurate similarity results. This analytical process yields the foundational data required for linked data construction.

Linked Data Technology

Linked data represents a collection of best practices that adopts the RDF data model, uses URIs (Uniform Resource Identifiers) to name data entities, and publishes and deploys instance and class data, thereby enabling data revelation and retrieval through HTTP protocols while emphasizing data interconnection and context information beneficial for human and machine understanding. As an ontology description method, its representation scope includes concepts, concept hierarchies, properties, property value types, relationships, relationship domain concept sets, and relationship range concept sets, with rules or axioms addable to represent more complex constraint relationships at the schema level. Linked data substantially follows these principles: RDF documents are named with Uniform Resource Locators (URIs); URIs must conform to Hypertext Transfer Protocol (HTTP); information pointed to by URIs must be provided in standard formats (RDF, SPARQL); and published information must contain URIs and other linked data principles.

Linked data serves as a crucial tool for constructing web knowledge organization systems. Its object identification and access mechanisms create favorable conditions for cross-regional information resource aggregation and traceability

while providing normative control for various object entities and numerous conceptual terms involved. Through standardized naming and referencing, linked data strictly limits data semantics and links to numerous related resource entities, where the "properties" of linked data are themselves resources. Linked data possesses domain-independent, machine-understandable characteristics that reduce resistance generated during data flow and transformation processes through human-machine or machine-machine interaction, better carrying semantic data for user access and machine processing.

Framework and Functions for Similar Document Aggregation in Academic Resource Websites Based on Linked Data

Academic resource websites selectively publish literature according to their disciplines. Limited by disciplinary hierarchy, coverage scope, and update frequency, these sites only perform simple classification or even uniform collection without categorization. Users must browse individually to locate needed documents. Consequently, effectively achieving similar document aggregation and accurately presenting retrieved similar document sets to users becomes critically important.

Aggregation Framework

The aggregation framework structure is shown in Figure 1 [Figure 1: see original paper]. This framework achieves transformation from loosely structured web information to linked data describing knowledge relationships, ultimately enabling linked document retrieval and recommendation through three units: (1) **Web Information Extraction Unit** extracts literature resources of different types, periods, and formats within academic resource websites, removing irrelevant content such as webpage code. (2) **Information Parsing and Document Association Calculation Unit** further extracts metadata and disciplinary knowledge from academic resource website documents, vectorizes them, calculates similarity between documents, establishes similarity graphs around each document, and creates conditions for linked data generation. (3) **Linked Data Generation and Retrieval Unit** uses academic documents obtained by the web information extraction unit and document similarity lists from the academic information parsing and document association calculation unit as foundations. Through a dynamic document generation system, it generates complete disciplinary network information linked data as core retrieval resources, seamlessly linking with user information retrieval behaviors to provide fast and accurate linked literature retrieval services, thereby promoting knowledge discovery and utilization in academic resource networks.

Aggregation Functions

Similar document aggregation in academic resource websites is implemented through three functions: web document collection and preprocessing, document

similarity calculation and filtering, and linked data generation and retrieval.

(1) Web Document Collection and Preprocessing
Most literature embedded in academic resource websites resides within webpage files and requires collection and preprocessing to form a foundational corpus for further analysis. The process is illustrated in Figure 2 [Figure 2: see original paper]. Academic resource website document collection includes webpage reading and structure parsing. Webpage reading essentially involves saving target webpages from the network to local storage through crawlers, offline browsers, or FTP tools. To display documents in browsers according to preset styles, both content and format are described through a complete set of markup tags, with each tag group forming a node. When extracting web document content, the page framework must first be parsed to accurately select nodes containing document content. Preprocessing involves cleaning collected webpage node content by removing redundant page code, format symbols, and other irrelevant elements, then saving the organized content by document unit for analysis.

(2) Document Similarity Calculation and Filtering
Similar document discovery in academic resource websites depends on calculating and filtering inter-document similarity. This study employs latent semantic analysis to compute semantic relationships among literature resources, building a latent semantic analysis vector space based on the document corpus from the web information extraction unit to calculate overall document similarity. Through hierarchical clustering of the document set, we determine a threshold for filtering document similarity to exclude low-similarity documents. After forming a new similar document list, we write it into linked data to achieve semantic aggregation of similar literature resources based on linked data. The document similarity calculation and filtering process is shown in Figure 3 [Figure 3: see original paper].

This process comprises two phases:
Phase 1: Building a latent semantic analysis vector space from the document corpus. The primary function involves tokenizing and removing stop words from raw content extracted from academic resource websites, establishing a document-term matrix, and calculating weights for term frequency and document relationships. This study uses Term Frequency-Inverse Document Frequency (TF-IDF) to compute document similarity weights. TF-IDF consists of two components: Term Frequency (TF) counts term occurrences in a document. To prevent weight fluctuations caused by significant frequency differences for certain key terms across documents, we apply logarithmic scaling to term weights for individual documents, compressing values to the 0-1 range. Since the document-term matrix is sparse, we add 1 before taking the logarithm of each term frequency: TF(t,d) = log(1 + f_{t,d}), preventing excessively small final TF-IDF values that would hinder comparison. Inverse Document Frequency (IDF) is the ratio of total documents in the collection to the number containing the current term: IDF(t,D) = log(N/|{d D, t d}|), representing term universality in the document collection. For document similarity judgment, terms

appearing in more documents within the collection have lower discriminative power and importance. TF-IDF is the product of TF and IDF: TFIDF(t,d,D) = TF(t,d) × IDF(t,D). The document-term matrix is built upon these weights adjusting the association degree between each document and vocabulary. By sequentially arranging all n documents under analysis and composing a matrix according to term frequencies from m terms contained in all literature, we form an n×m document-term matrix M.

Phase 2: Extracting high-similarity documents depends on overall similarity calculation and similar literature filtering. This study uses latent semantic analysis to compute the semantic structure of processed academic resource website documents. The document-term matrix M undergoes singular value decomposition to obtain term matrix, singular value matrix, and document matrix: M = UΣV^T. The product of the singular value matrix and document matrix ΣV^T represents the dimension-reduced document space vectors. Document similarity can be calculated using the cosine value between two dimension-reduced document vectors in the vector space; larger values indicate higher similarity between corresponding documents. After obtaining overall document similarity, we adopt a similarity threshold selection method based on document hierarchical clustering. Clustering analysis (or clustering) is the process of dividing data objects into subsets, where each subset is a cluster containing similar objects while objects from different clusters are dissimilar. Since it can partition large data collections into groups based on data similarity, it' s also called data segmentation. Hierarchical clustering treats each sample individual in the initial population as a separate class, uses Euclidean distance to evaluate similarity between classes, and merges the closest classes until clustering requirements are satisfied. Through hierarchical clustering of the document set, we obtain the maximum number of documents accommodated in a cluster, which serves as the basis for intercepting similar literature for each document in descending similarity order. Taking the median of the obtained similarity value set further excludes low-similarity documents, yielding a reasonable similarity threshold.

After these two phases, we exclude all low-similarity documents from the original similarity matrix, obtaining a document similarity matrix that records semantic similarity degrees among network literature of different types, formats, and cataloging rules.

(3) Linked Data Generation and Retrieval

The linked data generation and retrieval module writes the filtered similar document list into linked data using dynamic document technology. This linked data need not accommodate all literature from academic resource websites at once but can incrementally add new document information in real time. For specific types of literature resources, their unique semantics and semantic associations can be dynamically added to linked data, enabling customized expansion of core metadata ontologies to generate disciplinary linked data for particular subject categories.

Representation of Similar Resources in Academic Resource Websites

In linked data, representing relationships among academic resources through similar literature lists is an intuitive and retrieval-friendly method. We define similar literature resources as a subclass using the equivalence relationship (similarAs), generating corresponding similar literature lists for each document resource based on the document similarity matrix. Taking the blog post "From a Statistical Perspective on Deep Learning (2): Autoencoders and Free Energy" as an example, its similar literature list is shown in Figure 4 [Figure 4: see original paper].

Publishing academic resource network information resource semantic association lists in linked data format can intuitively display full-discipline literature association graphs, enabling easy access to entire disciplinary academic literature starting from linked data nodes. Users can access external related resources via Uniform Resource Locators and freely switch between different datasets. Since related literature is already sorted by relevance degree, simple queries can effectively reveal inter-resource relationships. Additionally, this approach enables semantic interoperability such as semantic retrieval.

Case Study: Real Academic Resource Website Data

Using the proposed ontology and linked data-based academic resource network similar literature aggregation framework, we constructed a demonstrative academic resource website similar literature aggregation system using R language to validate our approach.

Data and Preprocessing

We selected 78 editor-recommended documents published on the academic resource website "COS"(China of Statistics). Throughout the process, we used the XML package to extract literature links contained in the "Recommended Literature" column pages, pre-read the full-text pages of recommended documents through these links, extracted nodes containing document content, organized each document separately, and named each document using the final segment of its URI based on COS website link characteristics. We used the Rwordseg package (based on ICTCLAS segmentation software from the Institute of Computing Technology, Chinese Academy of Sciences, which employs a Hidden Markov Model) to uniformly read these documents for analysis, performing tokenization and stop word removal to reduce system resource consumption caused by meaningless words.

Similar Document Clustering and Discovery

We invoked the lsa package to read each tokenized document as a separate vector, forming a raw corpus. Due to differences between Chinese and Western languages, we adjusted the minimum word length to 1 to retain both Chinese

and Western terms while cleaning the corpus. We converted the cleaned corpus into a text matrix, which became a document vector set composed of tokenized terms. When calculating term frequency, we used TF-IDF as a method to balance high-frequency term weights, established the document-term matrix, performed singular value decomposition, and visualized the results as shown in Figure 5 [Figure 5: see original paper].

Figure 5 presents the dimension-reduced document-term matrix after singular value decomposition, clearly illustrating the relationships among document vectors formed by recommended documents published on the COS academic resource website. The concentrated cluster on the left demonstrates the semantic similarity of COS recommended literature, while eight outlier points scattered on the right indicate these eight documents are not closely associated with others. This phenomenon fully reflects that while academic resource websites primarily promote literature with high disciplinary thematic consistency, some gray literature with weak associations may also exist. Based on this, we calculated cosine similarity for the dimension-reduced document-term matrix to obtain the overall document similarity relationship matrix.

To accurately retrieve the most similar documents, we used a hierarchical clustering-based similarity threshold method. We invoked R' s proxy package to implement hierarchical clustering in two steps: calculating Euclidean distance between documents and performing hierarchical clustering using Ward' s method (sum of squared deviations). Ward' s method judges document clustering based on analysis of variance principles—if classification is reasonable, within-cluster sum of squares should be small while between-cluster sum of squares should be large. After calculation and visualization, we obtained the hierarchical clustering of COS recommended documents as shown in Figure 6 [Figure 6: see original paper].

Figure 6 displays the clustering of all recommended documents. When the height is set to 0.04, hierarchical clustering yields clusters containing 4 to 29 documents. To completely preserve high-similarity literature associations, we used the maximum cluster size as the initial value for extracting similar documents, then intercepted high-similarity data from each of the 29 documents. Summarizing these qualified similarity values and calculating their median yielded a similarity threshold of 0.6321.

Using this threshold, we further filtered the document similarity matrix data, retaining all similarity values above the threshold. If a document' s similarity with all other documents fell below the threshold, we selected only the single most similar document as its similar literature. Based on these inter-document associations, we generated a complete document set' s related literature list for each document in descending relevance order.

After generating the related literature lists, we used the dynamic document conversion tool rmarkdown to embed document metadata and related document lists into RDF/XML encoding segments to form complete linked data source

dynamic documents. Through simple format conversion, we obtained linked data recording similar documents. Using social network analysis and visualization tool iGraph, we visualized the document similarity relationships recorded in the linked data as shown in Figure 7 [Figure 7: see original paper].

In Figure 7, node size correlates with a document' s similarity to other documents—more similar documents produce larger nodes. Small nodes with single connections scattered at the network edges represent the few documents with low similarity to others (below threshold). Our similarity processing method maximally preserves these documents' relationships with others, enabling their discovery during similar literature retrieval. By publishing linked data online and applying the "Follow Your Nose" principle—where determining a URI pointing to some RDF allows loading the corresponding document—we can quickly obtain related literature retrieval results for any document in COS.

Linked data implementation offers high flexibility for similar literature recommendation: as the product of document analysis processes, user retrieval targets linked data directly, requiring only a simple query or inference to obtain results without comparing queries against each document sequentially, thereby improving retrieval efficiency. Retrieval results are provided directly by linked data in URI format, accessible through simple clicks that conform to general user habits—simple and convenient. Due to the independence of the analysis process, service interruption is unnecessary when documents change. The system can complete background analysis to generate new linked data that replaces old data without disturbing users.

Conclusion

This study employs latent semantic analysis to compute semantics contained in web documents, using the resulting similarity matrix as the foundation for generating linked data to demonstrate this method' s effectiveness in similar literature discovery. Our linked data construction is relatively simple, primarily characterizing inter-document similarity and utilizing similarity relationships to discover new similar literature. We have not yet classified and mined association rules for disciplinary knowledge involved in documents. Similar literature discovery based on objective knowledge systems and structures should better reflect disciplinary knowledge development trajectories and related literature association degrees. In subsequent research, we will introduce machine learning and other methods to perform deep aggregation of academic resource website literature content and knowledge.

References

[1] Zhang Yunzhong. From Integration to Aggregation: The Change of Digital Resources Re-organization Pattern in China [J]. Digital Library Forum, 2014(6): 16-20.

[2] Magerman T, Van Looy B, Song X. Exploring the Feasibility and Accuracy of Latent Semantic Analysis Based Text Mining Techniques to Detect Similarity Between Patent Documents and Scientific Publications [J]. Scientometrics, 2010, 82(2): 289-306.

[3] He Xiaoping, Li Di, Wang Mili, et al. A New Pre-Clustering-based Latent Semantic Analysis Algorithm for Document Retrieval[J]. Journal of Yunnan Nationalities University: Natural Sciences Edition, 2015, 24(3): 257-260.

[4] Wang W, Yu B. Text Categorization Based on Combination of Modified back Propagation Neural Network and Latent Semantic Analysis [J]. Neural Computing & Application, 2009, 18(8): 875-881.

[5] Olmos R, León J A, Jorge-Botana G, et al. New Algorithms Assessing Short Summaries in Expository Texts Using Latent Semantic Analysis [J]. Behavior Research Methods, 2009, 41(3): 944-950.

[6] Law J, Bauin S, Courtial J P, et al. Policy and the Mapping of Scientific Change: A Co-word Analysis of Research into Environmental Acidification [J]. Scientometrics, 1988, 14(3): 251-264.

[7] Tang Guoyuan, Zhang Wei. Development and Analysis of Subject Theme Evolution Based on Co-word Analysis Method [J]. Library and Information Service, 2015, 59(5): 128-136.

[8] Ren Jianhua, Shen Yanbin, Meng Xiangfu, et al. Document Clustering Based on Association Relations Between Terms[J/OL]. [2014-12-11]. Computer Engineering and Applications. http://www.cnki.net/kcms/detail/11.2127.TP.20141211.1528.053.html.

[9] Huang Xianying, Zhang Jinpeng, Liu Yingtao, et al. Short Text Similarity Algorithm Based on Term Mapping with Semantic[J]. Computer Engineering and Design, 2015, 36(6): 1514-1518, 1534.

[10] Xu Yong, Chen Jianguo, Hu Lingyun, et al. S&T Literature Hybrid Recommendation Algorithm Based on Generalized Semantic Similarity [J]. Information Studies: Theory & Application, 2013, 36(2): 96-99, 103.

[11] Wu Shufang, Liu Chang, Xu Jianmin. Research on Document Relevancy Based on Ontology Term Relations [J]. Journal of Modern Information, 2014, 34(9): 56-59, 176.

[12] Steyvers M, Griffith T. Probabilistic Topic Models[A]//Latent Semantic Analysis: A Road to Meaning [M]. Laurence Erlbaum, 2006.

[13] Landauer T K, Foltz P W, Laham D. An Introduction to Latent Semantic Analysis [J]. Discourse Processes, 1998, 25(2-3): 259-284.

[14] Leydesdorff L. Similarity Measures, Author Cocitation Analysis, and Information Theory [J]. Journal of the American Society for Information Science & Technology (JASIST), 2005, 56(7): 769-772.

[15] Structured Dynamic. Linked Data FAQ [EB/OL]. [2014-07-18]. http://structureddynamics.com/linked_{data}.html.

[16] Wang Haofen. Large-scale Knowledge Graph Technology [J]. Communications of the CCF, 2014, 10(3): 64-68.

[17] Berners-Lee T. Linked Data-Design Issues [EB/OL]. [2009-06-18]. http://www.w3.org/DesignIssues/LinkedData.html.

[18] Liu Wei. Overview on Linked Data: Concept, Technology and Implementation[J]. Journal of Academic Libraries, 2011, 29(2): 5-12.

[19] Han J, Kamber M, Pei J. Data Mining: Concept and Techniques[M]. Translated by Fan Ming, Meng Xiaofeng. 3rd Edition. Beijing: China Machine Press, 2012: 288-289.

[20] Tang X, Zhu P. Hierarchical Clustering Problems and Analysis of Fuzzy Proximity Relation on Granular Space[J]. IEEE Transactions on Fuzzy Systems, 2013, 21(5): 814-824.

[21] The R Project for Statistical Computing [EB/OL]. [2015-07-10]. https://www.R-project.org/.

[22] XML: Tools for Parsing and Generating XML Within R and S-Plus [EB/OL]. [2015-06-30]. http://CRAN.R-project.org/package=XML.

[23] Rwordseg: Chinese Word Segmentation[EB/OL]. [2013-12-15]. http://R-Forge.R-project.org/projects/rweibo/.

[24] lsa: Latent Semantic Analysis[EB/OL]. [2015-05-27]. http://CRAN.R-project.org/package=lsa.

[25] proxy: Distance and Similarity Measures[EB/OL]. [2015-07-08]. http://CRAN.R-project.org/package=proxy.

[26] rmarkdown: Dynamic Documents for R [EB/OL]. [2015-06-13]. http://CRAN.R-project.org/package=rmarkdown.

[27] Csardi G, Nepusz T. The iGraph Software Package for Complex Network Research[C]//Proceedings of InterJournal, Complex Systems. Cambridge, MA, USA. 2006: 1695.

[28] Antoniou G, Groth P, Hoekstra R, et al. A Semantic Web Primer [M]. Translated by Hu Wei, Cheng Gong, Huang Zhisheng. 3rd Edition. Beijing: China Machine Press, 2014.

Author Contributions Statement

Bi Qiang: Conceived research ideas and designed the study; Zhao Yiping: Conducted data collection, experiments, and drafted the manuscript; Bi Qiang and Zhao Yiping: Revised the final manuscript.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the online version of the journal at http://www.infotech.ac.cn:
[1] Zhao Yiping, Bi Qiang. category_{url}. Web links contained in the COS recommended article column.
[2] Zhao Yiping, Bi Qiang. article_{url}. All recommended article links extracted from the COS recommended article column.
[3] Zhao Yiping, Bi Qiang. ctl.csv. COS recommended article links and titles.
[4] Zhao Yiping, Bi Qiang. cate80.rar. Tokenized COS recommended articles.
[5] Zhao Yiping, Bi Qiang. ccsm.csv. Raw document similarity matrix.
[6] Zhao Yiping, Bi Qiang. cosArtSim.csv. Filtered document similarity matrix with low similarities removed.
[7] Zhao Yiping, Bi Qiang. cossim.rdf. COS similar recommended article linked data.

Received: August 13, 2015
Revised: October 4, 2015

Using Linked Data to Retrieve Similar Documents from the Academic Resource Websites

Zhao Yiping, Bi Qiang
(School of Management, Jilin University, Changchun 130022, China)

Abstract: [Objective] This paper studied the linked data from the Web, which is machine-readable, semantically meaningful and relationally descriptive. We examined these data' s effectiveness to improve the information organization of the academic resource websites (ARWs), with the purpose of retrieving more similar documents. [Methods] We first calculated the similarity of documents published in the ARWs with the help of the Latent Semantic Analysis (LSA) method. Then, chose documents with high similarities by the Hierarchical Cluster method, and created a document relation matrix. Finally, we used the dynamic document technology to generate a linked data index to search the ARWs. [Results] We built a preliminary ARWs linked data index, which helped us find similar documents more effectively from the ARWs. [Limitations] We investigated the similar documents retrieval technology from the perspective of statistical analysis. Therefore, further research is needed to locate similar documents from various subject areas with the support of deep learning technology. [Conclusions] We computed documents' similarity using LSA method to discover related documents of specific articles. The linked data could help us find more similar documents, while reducing the waiting time for similarity calculation.

Keywords: Linked Data; Latent Semantic Analysis(LSA); Academic Resource Websites(ARWs); Similarity

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