A Comparative Analysis of Domestic and International Research Hotspots on Linked Data: Postprint

Authors: Wang Zhenmeng

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

Abstract

[Purpose / Significance] Domestic Linked Data research remains in its early stages of development, with both theoretical and applied research yet to mature. This study provides reference for domestic Linked Data research through comparative analysis of hot research topics at home and abroad. [Method / Process] Through data analysis and literature review, comparative analysis of domestic and international research is conducted from the perspectives of publication volume, research communities, word frequency, and topic identification, thereby identifying existing gaps in Linked Data research between domestic and international contexts. [Results / Conclusion] Based on the comparative analysis, four development recommendations are proposed for domestic Linked Data research: research localization, emphasis on applied research, optimization of research community structure, and attention to disciplinary application of Linked Data.

Full Text

A Contrastive Analysis of Linked Data Research Hotspots at Home and Abroad

Wang Zhenmeng1,2

1University of Chinese Academy of Sciences, Beijing 100049
2Chengdu Documentation and Information Center, CAS, Chengdu 610041

Abstract

[Purpose/Significance] Domestic research on linked data remains in its early stages, with both theoretical and applied studies still immature. This contrastive analysis of research hotspots between China and abroad aims to

provide valuable references for domestic linked data research development. [**Method/Process**] Through data analysis and literature review, this study compares Chinese and international research from multiple perspectives, including publication volume, research communities, keyword frequency, and topic identification. [**Results/Conclusions**] Based on this comparative analysis, four recommendations are proposed to advance domestic linked data research: localization of research, emphasis on applied studies, optimization of research community structure, and focus on disciplinary applications of linked data.

Keywords: linked data; research hotspots; contrastive analysis; data analysis

Classification Number: G250.7

Citation Format: Wang Zhenmeng. A Contrastive Analysis of Linked Data Research Hotspots at Home and Abroad [J/OL]. Knowledge Management Forum, 2016, 1(3): 163-173 [Citation Date]. http://www.kmf.ac.cn/paperView?id=33.

1 Introduction

Since T. Berners-Lee first proposed the concept of linked data in 2006, researchers have embarked on in-depth investigations of this topic. Linked data represents a W3C-recommended specification for publishing and connecting diverse types of data, aiming to create a machine-readable web of data that describes everything in the world and their interrelationships. This vision seeks to build a semantically rich, interconnected knowledge ocean that enables people to accurately, efficiently, and reliably find, share, and utilize relevant information and knowledge on a large scale. Domestic attention to linked data began in 2008 when Zeng Lei, Liu Wei, and others formally introduced the concept at the "Frontier Issues of Libraries in the Digital Environment" seminar in Shanghai. Since then, domestic research on linked data has grown rapidly in volume, with research themes becoming increasingly broad. This paper employs co-word analysis and topic mining on relevant domestic and international research literature, utilizing tools such as SPSS and SATI for statistical and thematic cluster analysis. Through comparative analysis, this study attempts to identify the characteristics and shortcomings of domestic linked data research, providing references for its future development in China.

2 Data Sources and Research Methods

This study strives for precision and comprehensiveness in data collection. During the Chinese data collection process, the author identified additional keywords beyond the core term "linked data" and incorporated them into the search strategy. While ensuring comprehensive coverage, manual screening was further conducted on preliminary search results to account for polysemy in Chinese vocabulary, ultimately yielding 381 valid records. During the Chinese data screening process, the author found that these 381 records primarily belonged to

library and information science (G25) and computer science (TP3) fields, which was verified through institutional information of the first authors. In collecting English data, research direction constraints were applied to ensure effective comparison. The specific data source information is shown in .

Table 1 Chinese and English Data Source Information

Data Source	Chinese Data	English Data
Source Database	China Academic Journals Full-text Database	Web of Science Core Collection
Date	January 2, 2016	January 2, 2016
Search Strategy	time(2006,2016) and (subject (exact) = linked data or linked open data or 关联开放数据 or 链接数据 or 链接开放数据)	time(2006,2016) and (subject=linked data or linked open data) and (language=English) and (research direction=computer science or information science & library science)

This study employs three primary research methods: statistical analysis, keyword co-occurrence analysis, and comparative analysis. Statistical analysis reflects the current development status and future trends from a quantitative perspective. Keyword co-occurrence analysis reveals the relationship structure among research topics by examining the frequency with which keyword pairs appear together. Comparative analysis effectively highlights differences between comparison objects, with results providing valuable guidance for the disadvantaged party.

3 Data Analysis

Before formal data analysis, the author conducted basic cleaning of keywords in the acquired data, including merging Chinese-English equivalents, singular-plural forms, and eliminating low-value terms. To achieve deeper understanding of the data, the following sections conduct in-depth mining from perspectives of publication volume, keyword frequency, topic clustering, and core topic identification.

3.1 Publication Volume Analysis

Changes in publication volume can reflect a field's development status and future research trends. Statistical analysis of domestic and international linked data research by year is shown in [Figure 1: see original paper]. The concept of

linked data was first proposed in 2006 [3], entering a rapid development phase after 2008 and peaking in 2014, after which research enthusiasm declined. International linked data research demonstrates coordinated development of theory and practice, with the field making substantial progress as theoretical research themes diversified and deepened. In contrast, domestic research exhibits certain lag—while the concept was introduced in December 2008, the earliest published article was a translation by Yao Xiaole and Liu Wei of E. Summers et al.'s "LCSH, SKOS and Linked Data," which provided a basic introduction from the perspective of information organization [4]. Domestic linked data research has grown rapidly, showing an accelerating trend with a projected peak around 2017. However, influenced by declining international research interest, this peak may arrive earlier. Additionally, the author analyzed the disciplinary distribution: Chinese data shows approximately 51% computer science literature (202 articles) and 49% library and information science literature (196 articles), with 17 articles belonging to both fields. English data reveals approximately 92% computer science literature (1,275 articles) and about 8% library and information science literature (107 articles), with 68 articles spanning both disciplines.

3.2 Keyword Frequency Statistics Analysis

Keywords directly reflect article content, with frequency differences indicating varying importance. Statistical and co-occurrence analysis of domestic and international keywords reveals hotspot themes and their relationship structures. This study used SATI 3.2 for keyword extraction and frequency statistics, identifying 802 keywords in Chinese data and 2,420 in English data. Due to space limitations, only Chinese keywords with frequency $ $4 and English keywords with frequency $ 8*arelisted*(*see*), *excludingcorekeywords*"*linkeddata*"*and*"*linkedopendata*."*Tobetterillustraterelationshipsa*
frequencykeywords, *thispaperemploysbothkeywordco*−*occurrencematricesandkeywordsimilaritymatrices*.*C*
occurrencematricesintuitivelyshowpairwisefrequenciesandrelationships, *whilesimilaritymatricesnormaliz*
occurrencedatatoreduceerror, *withvaluesbetween*0−1*wherevaluescloserto*1*indicategreaterthematicsimilarit*
occurrencematricesforkeywordswithfrequency > 20(12×12*matrix*, *see*)*andpartialChinesesimilaritymatric*
10(11×$11 matrix, see).

3.3 Topic Clustering and Identification

Topic clustering is a common bibliometric method widely applied in domain analysis, utilizing word associations to form thematic clusters. This study conducted cluster analysis on Chinese keywords with frequency $ $5 (29 keywords, approximately 4% of total keywords), excluding core terms "linked data" and "linked open data." Based on previous analysis, the Ward clustering algorithm in SPSS was selected for hierarchical clustering, producing a dendrogram (see [Figure 2: see original paper]). The Ward algorithm was chosen because it emphasizes small within-cluster differences and large between-cluster differences, highlighting intra-cluster homogeneity—particularly suitable for thematic data samples. Using the Chinese data sample as reference, the top 4% of English key-

words (approximately 100 keywords with frequency $ $5) were selected. Due to SPSS limitations on data volume, CiteSpace was employed for English keyword clustering, yielding clustering results without manual adjustment (see [Figure 3: see original paper]).

The direct purpose of topic clustering is topic identification, whose accuracy depends heavily on clustering effectiveness. This phase identifies topics by tracing keywords back to original literature for manual identification, with results derived directly from the data. Since no strict hierarchical relationships exist among all keywords, and ignoring the hierarchical nature of dendrograms themselves, seven hotspot research themes were identified: linked data, ontologies and information organization (E), linked data publishing (C), linked data and library services (D), linked data knowledge discovery (F), linked data resource integration (G), linked data visualization (B), and linked data semantic interconnection (A).

English topic identification was based on 20 clusters formed by CiteSpace. After removing 8 isolated points without clear clustering, 12 thematic clusters emerged (see [Figure 3: see original paper]). Cluster positions reflect research theme locations, and since automatically generated cluster labels failed to clearly represent themes, manual backtracking of keywords within clusters produced new labels. presents these 12 thematic labels with corresponding keywords, noting that labels possess some subjectivity and are not entirely independent or equal, yet objectively reflect core research themes within clusters.

4 Comparative Analysis of Linked Data Hotspots at Home and Abroad

Comparative analysis effectively identifies differences between objects while guiding development for the disadvantaged party. Comparing Chinese and international linked data research reveals not only existing gaps but also future development directions for domestic research.

4.1 Differences in Research Communities

Significant differences exist between Chinese and international linked data research communities, which substantially contribute to thematic differences. These differences essentially represent varying research perspectives and methodologies, as researchers examine the same topic through their disciplinary lenses. Even overlapping disciplines naturally exhibit disciplinary characteristics in their research foci and outcomes, determined by researchers' knowledge structures and academic training. For linked data research, the primary disciplines are computer science and library/information science. The differences manifest primarily in the proportion of these disciplines: international library/information science literature accounts for 8% of the surveyed data, while domestic proportion reaches nearly 50%. From the institutional perspective, over 90% of domestic linked data researchers come

from library/information fields, compared to less than 10% internationally. Although researcher numbers, core authors, and individual publication counts could also reflect community differences, absolute quantity comparisons prove unreasonable when comparing domestic data against multiple countries. Overall, substantial differences exist in community composition, with domestic research dominated by library/information science researchers while international research primarily driven by computer science.

4.2 Differences in Research Theme Breadth

International linked data research themes are more comprehensive and complete than domestic ones, gradually forming a relatively complete research system, while domestic research still has numerous gaps. Comparing hotspot themes from section 3.3 reveals that international linked data research covers nearly all stages of the linked data lifecycle (see [Figure 4: see original paper]), whereas domestic research concentrates on early-stage phases, leaving later stages largely blank or just beginning. Comparing Chinese and international hotspot themes identifies domestic research gaps and emerging research points (see): themes that are current international hotspots but absent domestically represent research gaps, providing long-term development directions; themes that are international hotspots and have emerged but not yet developed into hotspots domestically represent emerging research points, offering references for current domestic research directions.

4.3 Differences in Research Theme Depth

Domestic linked data research from section 3.3 concentrates on seven themes. While international research on these issues is relatively mature or started earlier, significant differences exist even for identical topics, manifested in two aspects: diversity of sub-themes within the same theme, and quantity of applied research on the same theme. Taking "linked data publishing"—the first stage of the linked data lifecycle and a relatively mature domestic research theme—as an example, differences are shown in .

4.4 Differences in Development Trends and Research Models

Chinese and international linked data research exhibits not only current differences but also divergent future directions and trends. Domestic development will focus on emerging points and gaps such as linked data visualization and mining, with the primary approach remaining introductions of international research status and content. As domestic research deepens, applied studies across various themes may gradually increase. In contrast, international linked data quality research systems are more complete, with coordinated theoretical and applied research development having entered a slow-growth period, with research volume likely to further decline after 2015. Current bottlenecks in international research include linked data integration applications, disciplinary applications,

federated querying, and quality issues—these represent main development directions.

Overall, research models differ substantially. International research follows a progressive model, steadily advancing through coordinated theoretical and applied research to systematically address each lifecycle stage—a demand-driven approach. Domestic research, possibly due to late start and disciplinary background differences, heavily emphasizes introducing international research, primarily through reviews and case analyses, without following the lifecycle sequence. For instance, linked data integration has become a domestic hotspot while early-stage research remains immature.

5 Conclusion

Through analysis of publication volume, keyword frequency, themes, and research communities in domestic and international linked data research, this study not only identifies differences but also guides domestic development: First, optimize domestic research community structure by introducing new research forces—different disciplinary backgrounds determine research foci and forms, and introducing computer science researchers represents a pathway to deepen domestic linked data research and promote healthy development. Second, actively conduct applied research while deepening theoretical studies, focusing not only on concepts, principles, publishing methods, and integration theories, but also on developing and applying publishing tools, resource integration tools, and quality assessment tools. Third, attend to disciplinary applications of linked data, moving beyond library-centric entities to explore specialized applications in agriculture, environment, education, medicine, and other fields. Finally, emphasize localization of linked data research—current domestic research relies primarily on foreign-language resources, yet Chinese data transformation encounters unique issues such as URI configuration, making localization a prerequisite for developing domestic linked data initiatives.

References

[1] Liu Wei. Linked data: Concept, technology and application prospects [J]. Journal of Academic Libraries, 2011, 29(2): 5-12.
[2] Feng Lu, Leng Fuhai. Theoretical progress in co-word analysis methods [J]. Journal of Library Science in China, 2006, 32(2): 88-92.
[3] BERNERS-LEE T. Linked data [EB/OL]. [2016-03-06]. https://www.w3.org/DesignIssues/LinkedData.htm
[4] SUMMERS E, ISAAC A, REDDING C, et al. LCSH, SKOS and linked data [J]. Translated by Yao Xiaole, Liu Wei. New Technology of Library and Information Service, 2009(3): 8-14.
[5] SCHARFFE F, EUZENAT J. Linked data meets ontology matching: enhancing data linking through ontology alignments[C]//International conference on knowledge engineering and ontology development. Paris: Springer, 2011: 279-284.

[6] OSIMO D, MUREDDU F. Research challenge on opinion mining and sentiment analysis[J/OL]. [2016-03-20]. http://www.w3.org/2012/06/pmob/opinionmining.pdf.
[7] ZHAO J, MILES A, KLYNE G, et al. Linked data and provenance in biological data webs [J]. Briefings in bioinformatics, 2009, 10(2): 139-152.
[8] SPEISER S, HARTH A. Integrating linked data and services with linked data services [M]. Berlin: Springer, 2011: 170-184.
[9] HOGAN A, HARTH A, UMBRICH J, et al. Searching and browsing linked data with swse: the semantic web search engine [J]. Web semantics: science, services and agents on the world wide web, 2011, 9(4): 365-401.
[10] VRANDEČIĆ D, KRÖTZSCH M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.
[11] MENDES P N, MÜHLEISEN H, BIZER C. Sieve: linked data quality assessment and fusion[C]. Proceedings of the 2012 joint EDBT/ICDT workshops. Berlin: ACM, 2012: 1-6.
[12] BRUNETTI J M, AUER S, GARCÍA R, et al. Formal linked data visualization model[C]//Proceedings of international conference on information integration and web-based applications & services. Sydney: ACM, 2013: 97-104.
[13] PIEDRA N, CHICAIZA J A, LÓPEZ J, et al. An architecture based on linked data technologies for the integration and reuse of OER in MOOCs context [J]. Open praxis, 2014, 6(2): 171-187.
[14] ZAVERI A, KONTOKOSTAS D, SHERIF M A, et al. User-driven quality evaluation of dbpedia[C]//Proceedings of the 9th international conference on semantic systems. Graz: ACM, 2013: 97-104.
[15] CORRENDO G, SALVADORES M, MILLARD I, et al. SPARQL query rewriting for implementing data integration over linked data[C]//Proceedings of the 2010 EDBT/ICDT workshops. Berlin: ACM, 2010: 4.
[16] DI NOIA T, MIRIZZI R, OSTUNI V C, et al. Linked open data to support content-based recommender systems[C]//Proceedings of the 8th international conference on semantic systems. New York: ACM, 2012: 1-8.
[17] Fan Wei. Research on terminology services of controlled vocabularies [J]. Library and Information Service, 2012, 56(14): 34-39, 97.
[18] Liu Wei, Xia Cuijuan, Zhang Chunjing. Big data and linked data: The coming data technology revolution [J]. New Technology of Library and Information Service, 2013 (4): 2-9.
[19] Shen Zhihong, Li Jianhui, Zhang Xiaolin. Overview of linked data interconnection technology research: Applications, methods and frameworks [J]. Library and Information Service, 2013, 57(14): 5-11.
[20] Zhang Chunjing, Liu Wei, Xia Cuijuan. Open application protocol for linked data [J]. Journal of Library Science in China, 2012, (1): 43-48.
[21] Huang Yongwen. Review of linked data application research in libraries [J]. New Technology of Library and Information Service, 2010, 26(5): 1-7.
[22] Tang Yijie, Yang Rui, Liu Yi. Recent research hotspots and advances in SPARQL query technology [J]. Knowledge Management Forum, 2014(1): 1-6.
[23] Guo Shaoyou, Shang Yu, Wei Pengzheng, et al. Green linked data: Concept, requirements and engineering [J]. Library and Information Service,

2012, 56(23): 113-118.
[24] Tian Ye, Zhu Zhongming, Liu Shudong. Review of recommender systems based on linked data [J]. New Technology of Library and Information Service, 2013, 29(10): 1-7.
[25] Chen Tao, Xia Cuijuan, Liu Wei, et al. Research and implementation of linked data visualization technology [J]. Library and Information Service, 2015, 59(17): 113-119.
[26] Wu Yue, Li Zhanyu. Opening government data based on linked data [J]. Computer Knowledge and Technology, 2010 (11): 8688-8691.
[27] Wu Pengfei, Ma Fengjuan. Research and enlightenment on foreign linked data educational application projects [J]. e-Education Research, 2013, 34(4): 114-120.
[28] Wei Lai, Fu Yao. Linked data quality standards based on Green Linked Data [J]. Information and Documentation Services, 2013, 34(3): 69-72.
[29] Jiang Shiyin, Li Chunwang. Research on comprehensive evaluation model for scientific research institutions based on linked data [J]. Information Studies: Theory & Application, 2015, 38(11): 71-75.
[30] Ou Shiyan, Hu Shan, Zhang Shuai. Ontology and linked data-driven semantic integration method for library information resources and its evaluation [J]. Library and Information Service, 2014, 58(2): 5-13.
[31] Bai Haiyan, Liang Bing. Implementing semantic schema mapping between relational databases and linked data using D2R [J]. New Technology of Library and Information Service, 2011(8): 1-7.

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