Research on Data-Driven Mechanisms and Performance Optimization for Knowledge Discovery in Digital Libraries (Postprint)

Authors: Li Jie, Bi Qiang, Xu Pengcheng, Mou Dongmei

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] In a data-driven environment, exploring the data-driven mechanisms and optimization schemes of digital library knowledge discovery platforms provides theoretical support for their supply-side reform from a methodological and cognitive perspective. [Method/Process] By employing system dynamics methods, the dynamic formation mechanism of data-driven digital library knowledge discovery is presented through simulation; from a performance optimization perspective, granular computing methods are utilized to provide feasible solutions for its driving optimization. [Results/Conclusion] The data-driven factors influencing digital library knowledge discovery mainly include the data dimension, semantic association dimension, visualization dimension, and value dimension. From the perspective of dimension formation and performance interaction relationships, the data-driven nature of digital library knowledge discovery constitutes a dynamically evolving system with spiral development. The key point for its performance optimization lies in the degree of knowledge value exploitation from data. Empirical research demonstrates that employing knowledge granularity as the entry point for achieving optimization can effectively enhance the data-driven effectiveness of digital library knowledge discovery.

Full Text

Data-Driven Mechanism and Performance Optimization of Knowledge Discovery in Digital Libraries

Li Jie[1], Bi Qiang[1], Xu Pengcheng[1], Mou Dongmei[2]
[1]School of Management, Jilin University, Changchun 130022
[2]School of Public Health, Jilin University, Changchun 130021

Abstract:
[**Purpose/Significance**] Under the data-driven environment, exploring the data-driven mechanism and optimization scheme of knowledge discovery platforms in digital libraries provides theoretical support for supply-side reform from a methodological perspective. [**Method/Process**] Using system dynamics methods, this study presents the dynamic formation mechanism of data-driven knowledge discovery in digital libraries through simulation. From the perspective of performance optimization, granular computing methods are employed to provide feasible solutions for driving optimization. [**Result/Conclusion**] The data-driven factors influencing digital library knowledge discovery mainly include data dimension, semantic association dimension, visualization dimension, and value dimension. From the perspective of dimension formation and performance relationships, the data-driven knowledge discovery in digital libraries constitutes a dynamically spiraling development system. The key point of performance optimization lies in the exploitation degree of knowledge value within data. Empirical studies demonstrate that using knowledge granularity as the entry point for optimization can effectively enhance the data-driven effect of digital library knowledge discovery.

Classification Number: G250.76
Keywords: digital library, knowledge discovery, data-driven
DOI: 10.13266/j.issn.0252-3116.2019.03.001

---

2. Data-Driven Dimensions of Digital Library Knowledge Discovery

The knowledge discovery service process in digital libraries represents a transformative shift from data input to knowledge output, driven by the integrated forces of data fragmentation, semantic association, and knowledge visualization. Centered on fragmentation, the data dimension drive; oriented toward deep semantics and broad association, the semantic association dimension drive; focused on visual representation, meaning construction, and inference prediction, the visualization dimension drive; and measured by data flexibility and insight as standards for knowledge transformation, the value dimension drive collectively constitute the logical driver of digital library knowledge discovery. The specific driving dimensions and their involved elements and attributes are as follows:

2.1 Data Dimension

Digital library knowledge discovery proceeds from data directly to problems, achieving hierarchical deconstruction of different structures (structured, semi-structured, and unstructured) through fragmentation, semantic association, visualization, and knowledge capitalization, dedicated to the mining, parsing, creation, and domain application of data's transformation into knowledge capital. At the structured data-oriented level, the data dimension primarily involves:
Data quality, encompassing correctness and completeness of data sources, as

well as consistency, continuity, timeliness, precision, self-descriptiveness, and formalization degree, plus data accuracy and synchronization; **Data structure**, referring to the degree to which logical expression can be achieved through two-dimensional table structures—digital library literature knowledge discovery targets unstructured data, while database knowledge discovery targets structured data; **Data correlation**, concerning analytical approaches—data-driven digital library knowledge discovery primarily addresses problems through correlation analysis rather than causal analysis, where the degree and characteristics of data correlation play crucial roles in mining implicit knowledge and discovering patterns; **Data processing degree**, indicating preprocessing levels generally involving data auditing, cleaning, transformation, abstraction, integration, annotation, and sorting.

2.2 Semantic Association Dimension

Under data-driven conditions, digital library knowledge discovery services face the "semantic web" characterized by semantic analysis and association capabilities. The semanticization centered on semantic concepts, types, relationships, annotation, and reasoning, combined with the association drive focusing on association structure, strength, and rules, propels digital library knowledge discovery realization. The semantic association dimension under their cross-interaction mainly involves: **Semantic concepts**, conceptual units describing data with semantic attributes that aggregate to form semantic types; **Semantic relationships**, presenting intrinsic connections between data through logical ternary semantic roles and verb-core functions, where LSR (labeled semantic relations) can present both inter-concept relationships and relationship types; **Semantic annotation**, the process of associating entity concepts, attributes, and relationships with corresponding semantic descriptions through semantic metadata [?]; **Semantic reasoning**, a method applying specific meaning postulates within semantic system frameworks to reveal term semantic structures and relationships; **Association structure**, presenting internal/external association characteristics of data entities through data structure representation or specific methods; **Association relationships**, direct or indirect, implicit or explicit structured relationships between entities; **Association rules**, expressible as implication X→Y, where X and Y represent antecedent and consequent respectively, with support, confidence, and strength distinctions—strong association rules satisfy user-defined minimum support and confidence thresholds. Overall, semantic association dimension attributes involve semantic similarity, distance, relevance, overlap, matching, granularity, membership, radius, weight, proximity, tag breadth, link breadth, link depth, association dimension, strength, order, and granularity [?].

2.3 Visualization Dimension

Visualization-driven knowledge discovery represents both the presentation of data-knowledge network morphology and the process of knowledge analysis, prediction, and domain exploration. Through embedded visualization technologies, discovery platforms present knowledge domains with visual representation and intelligent characteristics to users, satisfying their retrieval, interpretation, and prediction tasks. This dimension involves: **Visual representation**, including surface forms (shape, color, texture) and deep forms (harmony, symmetry, balance, rhythm) presenting spatial relationships; **Meaning construction**, the visual thinking process where subjects develop new understandings based on visual perception, past experience, and evolving knowledge structures; **Knowledge network structure**, reflecting cooperation networks among authors, institutions, journals, and citation analysis networks of coupling, co-citation, and co-occurrence; **Knowledge network measurement**, indicators including centrality (node centrality, group centralization, closeness centrality, betweenness centrality), average path length, and cohesive subgroups; **Time series**, numerical sequences arranged chronologically to reflect statistical indicator changes, presenting entity development trajectories and aiding prediction—temporal ordering of data based on user spatial domains facilitates knowledge value development. Key attributes include visual metaphor degree, channel smoothness, visual salience, symbolization degree, k-core subnets, clustering coefficients, load degree, citation length, citation frequency, coupling degree, subject relevance, subject controllability, and subject knowledge structure.

2.4 Value Dimension

The data-driven function of digital library knowledge discovery aims to achieve a virtuous cycle in the "data→user→knowledge discovery" service space. As the direct performance-linked power source in this cycle, the value dimension concerns transformation capacity from data input to knowledge output, comprising: **Data flexibility**, the ability to transform data into data products (productization, commercialization); **Data insight**, the discoverable information value inherent in data, closely related to subjects' data awareness, experience, and analytical capabilities; **Data assets**, the transformation from data resources to assets enabling pricing, property rights attribution, and transactions [**?**].

---

3. Data-Driven Mechanism of Digital Library Knowledge Discovery

Digital library knowledge discovery platform data-driven emphasizes the decoupling of user needs and services in data-based problem-solving processes. Its power source derives from integrated driving of data dimension, semantic association dimension, visualization dimension, and value dimension elements. Analyzing the dynamic types of these dimensional elements clarifies the power formation mechanism.

From the perspective of these dimensional elements, knowledge discovery process drivers fall into two categories: internal power related to data, semantic association, and visualization themselves (endogenous power), and external power generated under external conditions (exogenous power). Based on system dynamics' common power types, this paper classifies digital library knowledge discovery data-driven dimension elements into endogenous and exogenous categories, as shown in Table 1 .

Table 1. Data-Driven Elements and Dynamic Types of Digital Library Knowledge Discovery

Dimension	Element	Endogenous/Exogenous • Dynamic
Data	Data quality	Endogenous
	Data structure	Endogenous
	Data correlation	Endogenous/Exogenous •
	Data processing degree	Exogenous •
Semantic Association	Semantic concept	Endogenous
	Semantic relationship	Endogenous
	Semantic annotation	Exogenous •
	Semantic reasoning	Exogenous •
	Association structure	Endogenous
	Association relationship	Endogenous
	Association rules	Exogenous •
Visualization	Visual representation	Endogenous
	Meaning construction	Exogenous •
	Knowledge network structure	Endogenous
	Knowledge network measurement	Endogenous/Exogenous •
	Time series	Exogenous •
Value	Data flexibility	Exogenous •
	Data insight	Exogenous •
	Data assets	Exogenous •

Note: • indicates the emphasized dynamic type

From the perspective of force manifestation, endogenous power represents direct driving through inherent forces, while exogenous power generates driving through external forces without directly acting on entities. As Table 1 shows, digital library knowledge discovery elements exhibit significant data-driven effects, with dimensions displaying non-single driving forms under different elements. Except for the value dimension, other dimensions serve as both endogenous and exogenous forces.

In the data dimension, data quality and structure act as endogenous forces; data correlation drives exogenously from an analytical thinking perspective but endogenously from a correlation degree perspective; data processing degree acts exogenously. In semantic association, semantic concepts, relationships, association structures, and relationships are endogenous; annotation, reasoning, and rules are exogenous. In visualization, visual representation and knowledge network structure are endogenous; meaning construction and time series are exogenous; knowledge network measurement drives both endogenously and exogenously. In the value dimension, influenced by data subjects, data flexibility, insight, and assets all drive exogenously.

Overall, digital library knowledge discovery endogenous power originates from data dimension and its interaction with semantic, association, and visualization factors, while exogenous power relates to technical factors of datafication, semantic-association-visualization, plus value elements linked to performance that indirectly drive knowledge discovery under subjective factor changes. Figure 1 [Figure 1: see original paper] illustrates the bidirectional causal relationships.

As shown in Figure 1, digital library knowledge discovery data-driven results from combined endogenous and exogenous forces. Data dimension-driven discovery primarily relies on endogenous power, with positive relationships among elements except data quality and structure, overall driving force being influenced by data quality, structure, processing degree, and correlation—particularly processing degree. Semantic association dimension-driven discovery experiences dual endogenous-exogenous influence, with semantic concepts, relationships, association relationships, and rules exerting greater driving force, showing positive relationships among most factors except semantic relationship usability and consistency. Visualization and value dimensions provide more exogenous power, with force magnitude significantly influenced by knowledge network structure, measurement, meaning construction, and data assets.

System dynamics' structured thinking permeates the entire driving system space. The data dimension forms the foundation for both force types, concerning data fragmentation realization. Semantic association and visualization dimensions serve as dual-driven technical middleware, jointly affecting knowledge conversion rates. The value dimension directly manifests performance effects, linking driving origins and destinations through feedback.

---

4. Performance Optimization of Digital Library Knowledge Discovery Data-Driven

4.1 Performance Optimization Path

A fundamental data science task involves exploring methodological tools for describing data structure properties. Combining information theory's definition of information with preliminary knowledge properties—bits plus semantics (attaching semantics to bit messages)—with knowledge representation conversion and structural conversion equivalence theories, and considering digital library knowledge discovery field space structural characteristics, reveals that data and knowledge structural property development is key to enhancing data-knowledge conversion rates. Therefore, performance optimization should adopt a structuralist perspective. In knowledge structural expression, knowledge structuring degree begins with knowledge granularity, where granular structure relates to knowledge architecture and can transform complex knowledge networks into simpler multi-level structures based on coupling properties. Using granular computing's multi-granularity hierarchical methods, this study achieves granular data-to-knowledge conversion through data granulation, optimizing digital library knowledge discovery performance through knowledge structure optimization. The specific optimization approach is shown in Figure 2 [Figure 2: see original paper].

4.2 Performance Optimization Method

Granular computing, integrating rough sets, fuzzy sets, quotient space, and cloud models, achieves complex problem simplification through multi-level, multi-granularity thinking. For digital library knowledge discovery field space structural constraints, granular computing based on granular structure aligns with data science's "data-based problem solving" trend, accelerating data-to-knowledge conversion efficiency through knowledge granulation structural performance improvement, thereby promoting performance optimization.

Multi-granularity analysis is an effective method in granular computing problem-solving, endowing data with semantic granularity attributes to advance granular data-to-knowledge semantic progression, accelerating knowledge granularity-based granular space construction, and enabling deep knowledge mining, discovery, and utilization [?]. Currently, China's mainstream digital library knowledge discovery platforms primarily serve research users with semantic retrieval services. Combining this reality, this study focuses on semantic retrieval decisions in discovery platforms for concrete data-driven performance optimization implementation, applying multi-granularity analysis based on multi-granularity rough set theory for service performance optimization. Following the principle of seeking common ground while reserving differences, performance discrimination and optimization are conducted from an optimistic multi-granularity fusion perspective [?].

In multi-granularity rough set data modeling, information system formal description resembles relational databases [?]. Let information system S=(U,AT,f), where $f_a:U\rightarrow V_a$ represents the relationship between U and V_a, meaning for any a AT, V_a is attribute a's value domain. For any x U, x's information vector is represented as [?]: f(x)={(a),f(a)_a|a AT}. In digital library knowledge discovery fields, platforms provide complete semantic retrieval decision support systems. Based on these definitions, let the user semantic retrieval decision system be:

DS={IS_i|IS_i=(U,AT_i,{V_a}*{a AT_i},f_i)},   X*{il} U/AT_j,   i=1,2…,m, l=1,2,…,t_i, Y_j U/{d}, j=1,2,…,k

Following optimistic multi-granularity rough set's core principle of seeking common ground while reserving differences (non-exclusive), using des(X_{il}) and dos(Y_j) for equivalent description of X_{il} and Y_j in DS, the multi-granularity fusion rule Z^o_{ij} can be defined as:

V_m i=1 des(X_{il}i)→des(Y_i), j=1,2,…,k

The rule's certainty Cer and support Supp are expressed as:

Cer(Z^o_{ij})=max_m |X_{il} Y_i|, Supp(Z^o_{ij})=max_m |X_{il} Y_i|

Based on difficulty comparing results due to local variable inconsistency, holistic thinking is commonly applied for judgment. Therefore, from a global decision perspective, this study uses overall certainty and overall support for system performance discrimination [?]. Using α°(DS) and β°(DS) to represent overall certainty and overall support respectively:

α°(DS)= Cer(Z$^{o_\{ij\})\times \mathrm{Supp}(Z}$o_{ij})
β°(DS)=1- Cer(Z$^{o_\{ij\})\times (1\text{-}\mathrm{Cer}(Z}$o_{ij}))

4.3 Performance Optimization Implementation

Targeting retrieval result performance optimization, this study randomly selected 7 student users of Jilin University's "Dingxin Chinese Discovery" as experimental subjects for user retrieval decision performance testing based on "granular structure" optimization theory. User professional backgrounds and retrieval content were unrestricted, with document type limited to journals and publication year limited to the recent five years.

Digital library knowledge discovery fields provide multiple retrieval channels—for example, Jilin University's platform offers different granularity-level channels including all fields, subject, abstract, title, and keywords in its advanced search window. For result output, sorting methods include publication date, collection, academic level, relevance, and citation count. For visualization, knowledge graph types include knowledge points, authors, and institutions based primarily on relevance.

This study used three different granularity-level retrieval channels (keywords, title, abstract) as conditional attribute sources, with relevance, citation count, and academic level as assessment factors under each source, and user satisfaction/dissatisfaction as performance attributes for "seeking common ground while reserving differences" multi-granularity retrieval decision performance testing.

The multi-source optimistic decision performance rule set is expressed as:

X_{11}→Y_1
X_{12}→Y_1
X_{21}→Y_2
X_{22}→Y_1
X_{23}→Y_2
X_{31}→Y_1
X_{32}→Y_1
X_{33}→Y_1
X_{34}→Y_1

Based on the granular space, we obtain:
G1={{e1,e3,e4,e5,e7},{e2,e6}}
G2={{e1,e2,e6},{e3,e4},{e5,e7}}
G3={{e3,e5,e6},{e1,e7}{e2},{e4}}

According to decision performance attributes, decision performance classes are: U{d}={{e2,e3,e5,e6}{e1,e4,e7}}

The multi-source optimistic decision performance rule set yields the RULL set shown in Table 3 .

Overall certainty is:
α°(DS)=0.0884

Overall support is:
β°(DS)=0.73016

Overall, optimistic multi-granularity decision shows good overall certainty and support. Using cross-validation and comparing decision rationality through clustering coefficient variations, the multi-granularity rough set decision performance rationality for user semantic retrieval in digital library knowledge discovery fields is shown in Figure 3 [Figure 3: see original paper].

As Figure 3 shows, considering cluster numbers from 1 to 6 for correctness prediction experiments, the analysis indicates that when K=2,3, improving driving performance through optimistic decision is most rational, with multi-granularity rough sets showing clear advantages in performance optimization and demonstrating good auxiliary effects for user semantic retrieval knowledge discovery services. Optimal performance occurs when the discrete bin number is 4 and cluster number is 3. For 7 users' retrieval needs, the knowledge discovery platform provides optimal retrieval performance when offering integrated keyword-title-abstract retrieval channels.

Single-granularity methods cannot adequately meet user semantic retrieval requirements. Multi-granularity methods provide more comprehensive and accurate decision support for knowledge discovery platform usage. Under multi-granularity optimistic decision, the system matches user decisions through "seeking common ground while reserving differences," enabling multi-channel retrieval decision matching and cross-fusion for different retrieval goals. Through comprehensive cross-retrieval of keywords, titles, and abstracts, the system provides optimized retrieval decision support, helping users find more comprehensive, accurate, and relevant knowledge information.

Empirical research on retrieval decision performance demonstrates that applying multi-granularity rough sets for digital library knowledge discovery data-driven performance optimization is effective. Multi-granularity semantic retrieval significantly improves data-to-knowledge conversion rates. During user semantic retrieval, multi-channel comprehensive retrieval better leverages digital library knowledge discovery platforms' knowledge mining, integration, and analysis capabilities, improving conversion efficiency from online collection data to user-demand-satisfying knowledge resources, enhancing data-driven service capabilities, and advancing the development from user fragmentation- and knowledge-demand-driven service recommendations to proactive multi-granularity hierarchical space knowledge service supply-side development.

In the big data environment, attention to data-intensive science is becoming a trend. For digital libraries, the key issue worth serious consideration is whether services can transform from extensive to precise models, opening new directions for digital library service transformation and achieving the shift from resource discovery to knowledge discovery [?]. The systematic force of data-driven digital library knowledge discovery services advances input data through joint driving of datafication, fragmentation, semanticization, association, and visualization toward intelligent knowledge. Viewing the causal relationships among data, semantic association, visualization, and value dimensions reveals that digital library knowledge discovery data-driven constitutes a spiraling dynamic ecosystem. The optimization ultimate goal is realizing knowledge value development of data resources. For value co-creation, data-driven optimization of digital library knowledge discovery, based on comprehensive consideration of factor performance relationships and using multi-granularity methods for knowledge structure optimization, can accelerate knowledge value conversion efficiency in the supply-side structural reform from datafication to intellectualization, providing more intelligent retrieval decision services for user semantic retrieval.

References

[1] WILLIAM SC. Data science: an action plan for expanding the technical areas of statistics[J]. International statistical review, 2010, 69(1):21-26.
[2] Li Jie, Bi Qiang, Zhang Han, et al. Research on data-driven knowledge discovery services in digital libraries[J]. Information and Documentation Services, 2018(4):6-14.
[3] TOLLE M, TANSLEY D, STEWART W, et al. The fourth paradigm: data-intensive scientific discovery[J]. Proceedings of the IEEE, 2011, 99(8):1334-1337.
[4] Chen Chen, Ma Xiaoting. Research on library user precise profiling based on small data[J]. Information and Documentation Services, 2018(5):57-61.
[5] Wang Shiwei. Era characteristics of data-driven and innovative transformation of library and information education[J]. Library and Information Science, 2016(1):15-20.
[6] Wang Dandan. Trends and practical cases of data-driven collection development[J]. Information and Documentation Services, 2016(5):74-78.
[7] POPOV B, KIRYAKOV A, KIRILOV A, et al. KIM-semantic annotation platform[M]. Berlin: Springer, 2003:834-849.
[8] Yan Jing, Bi Qiang, Li Jie. Construction of evaluation indicators for digital library resource aggregation quality[J]. Library and Information Service, 2017, 61(24):5-12.
[9] Chao Lemen. Data Science[M]. Beijing: Tsinghua University Press, 2016.
[10] Liu Shijun, Lu Xudong, Cui Lizhen. The Way to Win for Data-driven Enterprises—Data-Driven Innovation[M]. Beijing: Publishing House of Electronics Industry, 2017.
[11] Wang Guoyin, Zhang Qinghua, Hu Jun. Survey on granular computing re-

search[J]. CAAI Transactions on Intelligent Systems, 2007(6):8-26.
[12] LIN T. Data mining and machine oriented modeling: a granular computing approach[J]. Applied intelligence, 2000, 13(2):113-124.
[13] Wang Guoyin, Zhang Qinghua. Research on rough set uncertainty under different knowledge granularities[J]. Chinese Journal of Computers, 2008(9):1588-1598.
[14] Lin Guoping. Research on multi-granularity rough computing theory and methods[D]. Taiyuan: Shanxi University, 2016.
[15] Bi Qiang, Yan Jing, Li Jie. New situations and requirements for digital library service transformation in the big data era[J]. Information Studies: Theory & Application, 2017, 40(12):12-16,5.

Author Contributions:
Li Jie: Proposed research proposition and framework, wrote the full text;
Bi Qiang: Proposed research framework;
Xu Pengcheng: Revised the paper;
Mou Dongmei: Oversaw outline.

Received: 2018-07-08
Revised: 2018-10-10
Pages: 6-13
Responsible Editor: Yi Fei

Research on Data Driven Mechanism and Performance Optimization of Knowledge Discovery in Digital Library

Li Jie[1], Bi Qiang[1], Xu Pengcheng[1], Mou Dongmei[2]
[1]School of Management, Jilin University, Changchun 130022
[2]School of Public Health, Jilin University, Changchun 130021

Abstract: [Purpose/significance] Under the data-driven environment, exploring the data-driven mechanism and optimization scheme of knowledge discovery platforms in digital libraries provides theoretical support for supply-side reform from the perspective of methodology. [Method/process] By means of the system dynamics method, the data-driven dynamic formation mechanism of digital library knowledge discovery is presented through simulation. From the perspective of performance optimization, the granular computing method is used to provide a feasible solution for its drive optimization. [Result/conclusion] The data driving factors that influence the knowledge discovery of digital library mainly include data dimension, semantic association dimension, visualization dimension and value dimension. From the perspective of the formation of dimensions and the role of performance, the data drive of digital library knowledge discovery is a dynamic system of spiral development. The key point of performance optimization lies in the exploitation degree of knowledge value of data. According to the experimental studies, the knowledge granularity as the starting point to achieve its optimization can better improve the data-driven effect of digital library knowledge discovery.

Keywords: digital library; knowledge discovery; data-driven

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