Digital Library Numerical Knowledge Element Retrieval System Design Postprint

Authors: Huang Rong, He Yangyuqi, Wang Zhongyi, Li Chunya

Date: 2023-08-27T00:00:00+00:00

Abstract

[Purpose/Significance] To satisfy the personalized retrieval demands of digital library users for numerical knowledge and provide them with fine-grained knowledge services. [Method/Process] Based on in-depth analysis of numerical knowledge elements, this paper proposes methods for identifying, extracting, indexing, and retrieving numerical knowledge elements in digital libraries, and constructs a retrieval system oriented towards numerical knowledge elements. [Results/Conclusion] Case analysis demonstrates that fine-grained knowledge services based on numerical knowledge elements can improve the efficiency of retrieving and utilizing numerical knowledge and enhance user satisfaction to a certain extent.

Full Text

Design of a Numerical Knowledge Element Retrieval System for Digital Libraries

Rong Huang[1], Yangyuqi He[1], Zhongyi Wang[1], Chunya Li[2] [1]School of Information Management, Central China Normal University, Wuhan 430079 [2]School of Business, Nantong Institute of Technology, Nantong 226002

Abstract:

[**Purpose/Significance**] To meet personalized retrieval needs for numerical knowledge among digital library users and provide fine-grained knowledge services. [**Method/Process**] Based on in-depth analysis of numerical knowledge elements, this paper proposes methods for identifying, extracting, indexing, and retrieving numerical knowledge elements in digital libraries, and constructs a retrieval system oriented toward numerical knowledge elements. [**Result/Conclusion**] Case analysis demonstrates that fine-grained knowledge services based on numerical knowledge elements can improve the efficiency of re-

trieving and utilizing numerical knowledge and enhance user satisfaction to a certain extent.

Keywords: numerical knowledge element; knowledge element identification; knowledge element indexing; knowledge element retrieval
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2018.14.015

Since the 21st century, with the development of science and technology represented by information technology, people have entered an era of information explosion, and the emergence of knowledge services has brought many challenges to digital libraries [**?**]. Most existing digital library products adopt a subject-term-based model for resource organization and services, where the basic unit of knowledge service remains the document, making it impossible to provide fine-grained knowledge services for specific problems [**?**]. Compared with document-level knowledge units, people increasingly hope to directly retrieve knowledge points of interest, which requires digital libraries to gradually deepen the control unit of knowledge from coarse-grained document units to fine-grained knowledge element units [**?**], achieving a shift from managing knowledge carriers and attribute features to managing knowledge content itself—that is, transforming indirect knowledge management methods into direct ones [**?**]. A knowledge element is an indivisible knowledge unit with complete knowledge expression [**?**]. Based on content differences, knowledge elements can be categorized into theoretical and methodological knowledge elements, numerical knowledge elements, factual knowledge elements, and other types [**?**]. Numerical knowledge elements refer to knowledge units that exist in numerical form and describe numerical attributes (such as time, length, height, weight, percentage, sales, profit, etc.) of objective things or events [**?**]. Numerical knowledge elements are crucial for promoting the utilization of numerical knowledge, improving retrieval efficiency, and helping people discover potential and implicit numerical relationships. While most current scholars study numerical knowledge elements from theoretical perspectives [**?**], more effective methods for extracting complete and accurate numerical knowledge elements from text require further research. Therefore, this study takes digital library collections as the research object, investigating the identification, extraction, indexing, and retrieval of numerical knowledge elements in digital libraries to refine the granularity of knowledge services and improve service efficiency.

2 Identification and Extraction of Numerical Knowledge Elements

2.1 Identification and Extraction Rules for Numerical Knowledge Elements

Identifying numerical knowledge elements from digital library collections should first consider the existence forms of knowledge resources [**?**]. Knowledge is stored not only in traditional literature databases but also widely distributed in patent

data, industry standards, scientific and technical reports, and other specialized resource databases. This study focuses on numerical knowledge elements, which are mostly described at the sentence level in digital collections, making them suitable for rule- and pattern-based identification methods. Therefore, determining what types of numerical knowledge elements exist and how to construct identification rules are key to identifying numerical knowledge elements from digital collections.

Based on functional roles, numerical values can be divided into three categories: basic numerical values, process numerical values, and result numerical values. Consequently, this study classifies numerical knowledge elements in digital collections into basic numerical knowledge elements, process numerical knowledge elements, and result numerical knowledge elements. Different types have different description patterns, sentence structures, and complexity levels. While numerical knowledge element processes can sometimes be determined through individual values, description using sentence groups or paragraphs is more complete and accurate.

To induce identification rules for these three types, this study employed text analysis methods. First, 20 articles were selected from core journals across 13 disciplines, totaling 260 documents. These papers were segmented into sentences, and complete sentences containing numerical information were extracted. After screening and analysis, common expression patterns for different types were identified, forming the description rules shown in Table 1 .

Table 1 Identification Rules for Numerical Knowledge Elements

Numerical Knowledge Element Structure	Numerical Knowledge Element Identification Rules	Examples
Basic Numerical Knowledge Element	Time + Subject + Source + Predicate + (…year…month…day ~ /to…year…month…day)/(…year…month…day of…/by…/date/time is…) + Subject + (from/in/with/select…) + Source + (retrieve/collect/gather/distribute/obtain/select/download/provide/conduct/get + Numerical Value + Unit (e.g., items, papers, pieces, yuan, etc.) + Indicator	By December 31, 2010, 100 relevant papers were downloaded from Chinese citation database

Numerical Knowledge Element Structure	Numerical Knowledge Element Identification Rules	Examples
Process Numerical Knowledge Element	Time + Subject + Indicator + Predicate + Numerical Value + Unit	On May 8, 2010, displacement ductility coefficients of all specimens reached 3.0
Result Numerical Knowledge Element	Time + Subject + Indicator + Predicate + Numerical Value + Unit	By July 20, 2016, thermal weight loss rate within 300-600℃ range

2.2 Rule-Based Identification and Extraction of Numerical Knowledge Elements

Based on the structure and identification rules of numerical knowledge elements, this study designed an extraction method (Figure 1 [Figure 1: see original paper]). The basic flow includes: text segmentation, word segmentation and POS tagging, sentence filtering, numerical knowledge element attribute identification and extraction.

Figure 1 Extraction Process of Numerical Knowledge Elements

2.2.1 Text Segmentation Selected literature resources (PDF format) are converted to plain text, removing irrelevant information such as bibliographies and images. Text is then segmented based on sentence delimiters (e.g., 。; ? !

etc.).

2.2.2 Word Segmentation and POS Tagging Sentences are processed to generate phrase structures with POS tagging, assigning each word its most likely category (noun, verb, numeral, measure word, etc.). Stop words with minimal semantic content (adjectives, articles, etc.) are removed.

2.2.3 Matching The identification rules define clue words (numerical values, units, etc.) and their collocation structures for different numerical knowledge element types. Based on these rules, segmented sentences are matched to locate sentence groups or paragraphs containing numerical knowledge elements, forming a candidate sentence set for extraction.

2.2.4 Numerical Knowledge Element Extraction This step extracts attributes: subject, indicator, time, predicate, numerical value, unit, and source.

1. **Subject Identification**: Subjects mainly include region, industry, and institution types, identified using industry lexicons, institution feature word lists, and geographical dictionaries. If no region is specified, "China" is used by default.
2. **Indicator Identification**: Indicators are the numerical information themes, typically nouns adjacent to numerical values or units. Chinese word segmentation and POS tagging extract indicators, with an indicator library assisting subject extraction.
3. **Time Identification**: Time expressions are complex, including compound time phrases and duration words (e.g., "by June this year," "morning of October 5, 2015"). Four generalized patterns are used: time (e.g., 9:40), date (e.g., April 12, 2017), time words (e.g., this year, morning), and duration (e.g., one month, two years). The time nearest to and most recent before the indicator in the same sentence is selected. Elements without specific time are discarded.
4. **Predicate Identification**: Predicates typically precede numerical values or indicators as verbs or verb-preposition combinations (e.g., "decreased compared to last year").
5. **Numerical Value Identification**: Based on Chinese usage, numerical information has three categories: cardinal numbers (integers, decimals, fractions, etc.), ordinal numbers (combining cardinal numbers with "第"), and special numerals (using non-cardinal characters to express quantity, degree, or range, e.g., several, most).
6. **Unit Identification**: Units are measure words combining with numerical values (e.g., items, pieces, yuan), extracted using finite state machine algorithms based on measure word patterns.
7. **Source Identification**: The source document and its URL are identified during text analysis.

2.2.5 Numerical Knowledge Element Generation Results are stored to form a numerical knowledge element base for sharing and analysis.

3 Indexing and Retrieval of Numerical Knowledge Elements

Numerical knowledge elements are the basic building blocks of numerical knowledge construction in digital libraries. Their indexing and retrieval are crucial for storage, retrieval, and utilization.

3.1 Description Framework for Numerical Knowledge Elements

Current domestic research on numerical knowledge element description frameworks is limited and varied. This study proposes a more generalized entity object structure framework from three levels: knowledge identification, knowledge description, and knowledge relation (Table 2).

Table 2 Description Framework for Numerical Knowledge Elements

Level	Components
Knowledge Identification	Numerical knowledge element name (high-level summary of content)
Knowledge Description	Time, subject, indicator, predicate, numerical value, unit
Knowledge Relation	Source (links to carrier containing the element)

3.2 Indexing of Numerical Knowledge Elements

Indexing involves constructing indexes on discussed topics and attributes (numerical value, subject, indicator, etc.) to enable fast and accurate retrieval [**?**]. Based on the description framework, this study indexes from three aspects: knowledge identification, knowledge description, and knowledge relation. The indexing flow is shown in Figure 2 [Figure 2: see original paper], including information extraction, word segmentation, feature extraction, and index creation modules.

Figure 2 Indexing Process of Numerical Knowledge Elements

3.2.1 Information Extraction This module extracts information from the numerical knowledge element base to build indexes. Using the high-performance Lucene search engine architecture, the index structure uses Document as the basic unit, which consists of multiple fields. Information extraction transforms into extracting fields from numerical knowledge elements. Since all facets (name, time, subject, indicator, predicate, numerical value, unit, source) must be displayed to users, they are all extracted as Document fields (Figure 3 [Figure 3: see original paper]).

Figure 3 Document Creation

Different Field types have different requirements:

1. **Keyword Field**: Not analyzed but indexed verbatim and stored. Suitable for original values that must be preserved entirely. The "time," "numerical value," and "unit" fields are defined as Keyword type.
2. **UnIndexed Field**: Neither analyzed nor indexed, but stored for display with search results. The "source" field is defined as UnIndexed type.
3. **Text Field**: Analyzed and indexed, optionally stored. The "name," "subject," "indicator," and "predicate" fields require analysis and indexing, thus defined as Text type.

3.2.2 Text Segmentation This study uses the Stanford Segmenter for Chinese word segmentation and POS tagging, supporting user-defined dictionaries. A domain-specific dictionary was created to handle professional terminology in numerical knowledge elements, improving segmentation accuracy.

3.2.3 Index Creation This module creates inverted indexes for numerical knowledge element instances, generating seven index files: name, time, subject, indicator, predicate, numerical value, and unit.

1. **Document Generation**: All fields are stored in a Vector array for Lucene traversal. Sample code:

```
Document doc = new Document();
Field f1 = new Field("name", "value1", Field.Store.YES, Field.Index.TOKENIZED);
Field f2 = new Field("time", "value2", Field.Store.YES, Field.Index.UN_{TOKENIZED});
...
doc.add(f1); doc.add(f2); ...
```

2. **IndexWriter Initialization**: Creates an indexer to add Documents, merge index segments, and control index operations.
3. **Index Creation**: Using `addDocument(Document doc)` to add Documents to the index directory.
4. **Index Optimization**: The `optimize()` method merges all index files into one, reducing file count and improving retrieval speed.

3.3 Retrieval of Numerical Knowledge Elements

3.3.1 Name-Based Retrieval Users often cannot precisely locate complete names due to insufficient experience or query processing limitations, leading to low recall and precision. Since numerical knowledge element names combine multiple keywords, this study employs fuzzy retrieval. Fuzzy retrieval sets membership degree v [0,1] for query term x in documents, where higher v indicates greater relevance [**?**]. For example, querying "name=information" retrieves elements containing "information" in names, such as "China Information Industry" or "Industrial Information Department."

3.3.2 Boolean Logic Retrieval The description group contains six attributes: subject, indicator, time, predicate, numerical value, and unit. To improve recall and precision, Boolean logic retrieval is used, employing binary variables corresponding to knowledge element features and extracted query terms [**?**]. Queries can express Boolean relationships (and, not, or) among retrieval terms.

4 System Implementation

4.1 Development Tools and Environment

Storage: MySQL 5. Development: Eclipse Mar 2 for indexing and retrieval functions. Environment: Windows 10 Enterprise, Java J2EE 1.7, Tomcat 6.045.

4.2 System Implementation

Economic-themed literature was downloaded from CNKI. Numerical knowledge elements were extracted and stored using the proposed methods, and a search engine was implemented. The retrieval interface is shown in Figure 4 [Figure 4: see original paper].

Figure 4 Numerical Knowledge Element Retrieval Interface

4.2.1 Name Retrieval Fuzzy retrieval is implemented for name queries. For example, searching "information industry" retrieves elements with "information industry" in their names (Figure 5 [Figure 5: see original paper]).

Figure 5 Name Retrieval Results Interface

4.2.2 Boolean Logic Retrieval Multiple fields can be combined. For example, "subject=China" AND "time=2015" OR retrieves elements where subject is "China" or time is "2015" (Figure 6 [Figure 6: see original paper]).

Figure 6 Boolean Logic Retrieval Results Interface

4.3 Experimental Evaluation

A task-oriented subjective evaluation method assessed effectiveness and practicality. Thirty digital library users (15 undergraduates, 15 postgraduates) completed four retrieval tasks (Table 3) using three reference systems (Baidu Scholar, CNKI, Baidu Zhidao) and the numerical knowledge element search engine. Mouse clicks were recorded.

Table 3 Retrieval Tasks

Task	Description
Q1	Which country's GDP reached 10.87 trillion USD in 2015?
Q2	In which year did China's GDP reach 10.87 trillion USD?

Task	Description
Q3	What was China's GDP in 2015?
Q4	What indicator of China reached 10.87 trillion USD in 2015?

After tasks, users completed an experience questionnaire (Table 4) using a 5-point Likert scale (1=very dissatisfied to 5=very satisfied).

Table 4 User Experience Questionnaire

Evaluation Item	Score
Numerical Knowledge Element Search Engine	

Normalized satisfaction scores (A) were calculated using formula (1), where i=task number, j=user number, q_{ij}=user j's score for task i:

$$A_i = \frac{\sum_{j=1}^{30} q_{ij}}{5 \times 30}$$

Average clicks (B) were calculated using formula (2), where p_{ij}=user j's clicks for task i:

$$B_i = \frac{\sum_{j=1}^{30} p_{ij}}{30}$$

Figure 7 [Figure 7: see original paper] shows performance across four systems.

Results: - **User Experience**: The numerical knowledge element search engine scored highest (0.79, 0.91, 0.86, 0.83), indicating highest satisfaction. CNKI and Baidu Zhidao scored 0.5-0.7 (moderate satisfaction). Baidu Scholar scored lowest (0.37, 0.32, 0.24, 0.23). - **Clicks**: Baidu Scholar required most clicks (average 8), followed by CNKI (6) and Baidu Zhidao (4). The numerical knowledge element search engine required fewest clicks (average 2).

Analysis: 1. Baidu Scholar provides only knowledge carrier clues (title, abstract, author), requiring users to access documents and manually locate information, increasing cognitive load. 2. CNKI provides carrier clues but allows direct access to documents, resulting in slightly fewer clicks and higher satisfaction than Baidu Scholar. 3. Baidu Zhidao directly provides knowledge content with low cognitive cost, achieving second-highest satisfaction. 4. The numerical knowledge element search engine outperforms Baidu Zhidao because its resources come from digital libraries (higher quality), offers specialized retrieval entry points (time, unit), and operates on fine-grained knowledge elements, requiring minimal clicks and delivering highest satisfaction.

Conclusion

This study addresses digital resource management in digital libraries by proposing identification methods and extraction processes for numerical knowledge elements, developing a description framework and indexing/retrieval methods, and implementing a search engine that demonstrates feasibility. Future work will focus on knowledge element linking relationships to build a complete networked knowledge element system, improving digital library knowledge services.

References

[?] Wang Xinjun, Wang Haixin. Reflections on library knowledge services under big data background [?]. Library Work and Study, 2014(11): 75-78. [?] Zhao Junna. Research on subject services in university libraries for scientific research [?]. Hefei: Anhui University, 2014. [?] Xue Tiao. Visual analysis of evolution path, research hotspots and frontiers in domestic library subject knowledge service field [?]. Library and Information Service, 2012, 56(15): 9-14. [?] Zhang Haitao, Song Tuo, Liu Jian. Research on one-stop knowledge service model of university libraries [?]. Information Science, 2014(6): 104-108. [?] Bi Chongwu, Wang Zhongyi, Song Hongwen. Research on multi-granularity integrated knowledge service of digital library based on knowledge elements [?]. Library and Information Service, 2017, 61(4): 115-122. [?] Wen Tingxiao. Research on evolution and evaluation of knowledge units [?]. Library and Information Service, 2007, 51(10): 72-76. [?] Ronald M. Knowledge management systems: information and communication technologies for knowledge management (third edition) [?]. Berlin: Springer, 2007: 265-278. [?] Yuan Xiaoling. Knowledge indexing based on knowledge elements [?]. Library Science Research, 2007(6): 45-47. [?] Xiao Hong, Xue Dejun. Research on numerical knowledge element mining based on large-scale real texts [?]. Computer Engineering and Applications, 2008, 44(30): 150-152. [?] Fu Lei. Design and implementation of knowledge element indexing system [?]. Wuhan: Central China Normal University, 2005: 32-35. [?] Wu Chao, Zheng Yanning, Hua Bolin. Review of research progress on numerical information extraction [?]. Journal of Library Science in China, 2014, 40(2): 107-119. [?] Wen Youkui. Knowledge element mining [?]. Xi'an: Xidian University Press, 2009. [?] Jiang Yongchang, Yang Hongyan, Zhang Libo. Research on knowledge organization based on knowledge elements and its system service functions [?]. Information Studies: Theory & Application, 2007, 30(1): 37-40. [?] Yuan Yang, Xiao Hong. Knowledge service optimization based on automatic editing of knowledge element library [?]. Science-Technology & Publication, 2017(6): 22-25. [?] Robert LP, Ahuja MK. Social capital and knowledge integration in digitally enabled teams [?]. Information Systems Research, 2008, 19(3): 314-334. [?] Taha A. Networked library services in a research-intensive university [?]. Electronic Library, 2012(6): 844-856. [?] Hua Bolin. Research on types and description rules of method knowledge elements in academic papers [?]. Journal of Library Science in China, 2016, 42(1): 30-40. [?] Li Jing. Research on knowledge services based on knowledge organization [?]. Tianjin:

Tianjin Normal University, 2008. [?] Liu Chunyong. Research on information retrieval model in Chinese question answering system [?]. Chongqing: Chongqing University, 2007. [?] Ji Yongzheng. Analysis of search engine retrieval technology and retrieval effectiveness [?]. Journal of Qinghai University, 2006, 24(6): 98-100. [?] Drucker. Managing in the next society [?]. Translated by Zhao Gancheng. Shanghai: Shanghai Translation Publishing House, 1999: 211.

Author Contributions: - Rong Huang: Paper architecture design and writing - Yangyuqi He: Paper proofreading and editing - Zhongyi Wang: Paper concept design and revision - Chunya Li: Empirical design and results analysis

---

Abstract: [Purpose/significance] This paper aims to meet personalized retrieval needs of digital library users for numerical knowledge and provide fine-grained knowledge services. [Method/process] Based on analysis of numerical knowledge elements, it proposes methods for identifying, extracting, indexing, and retrieving numerical knowledge elements in digital libraries, and constructs a retrieval system for numerical knowledge elements. [Result/conclusion] Case study shows that fine-grained knowledge services based on numerical knowledge elements can improve efficiency and user satisfaction in retrieving and using numerical knowledge.

Keywords: numerical knowledge; knowledge element identification; knowledge element indexing; knowledge element retrieval

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