Decision Tree-Based Multi-Source Bibliographic Metadata Fusion Research Postprint

Authors: Li Jing, Hu Qian, Li Xiang, Xiao Bing

Date: 2023-04-01T15:51:24+00:00

Abstract

[Purpose/Significance] Constructing a multi-source literature metadata fusion model contributes to enhancing the overall quality of literature metadata, facilitating metadata management and utilization in resource discovery systems, and optimizing user experience in resource discovery services. This study optimizes the previously proposed literature metadata deduplication strategy by transitioning from an experience-based approach to an automated one, thereby increasing the automation level of the entire process while ensuring the effectiveness of both deduplication and fusion. [Method/Process] To address the issue that metadata fields vary across different literature types and across different sources for the same literature, which necessitates distinct deduplication methods, this paper proposes an automated decision tree-based multi-source literature metadata fusion model that transforms the deduplication problem into a classification problem. Features are selected and a decision tree is constructed based on feature similarity, upon which metadata deduplication and fusion are implemented. Experiments are conducted using metadata from various types of literature resources to validate the effectiveness of the proposed strategy. [Results/Conclusion] The results demonstrate that for metadata across five literature types, the deduplication strategy achieves an accuracy of over 99% and a recall rate of over 98%, indicating favorable overall performance. Regarding the effectiveness of the fusion strategy, the quality improvement ratios of metadata fields for patents, dissertations, journal articles, conference papers, and books are 15.15%, 36.80%, 15.29%, 52.63%, and 15.38% respectively, all exhibiting significant enhancement.

Full Text

Research on Multi-Source Document Metadata Fusion Based on Decision Trees

Li Jing, Hu Qian, Li Xiang, Xiao Bing School of Information Management, Central China Normal University, Wuhan 430079

Abstract

[**Purpose/Significance**] Constructing a multi-source document metadata fusion model helps improve the overall quality of document metadata, promotes metadata management and utilization in resource discovery systems, and optimizes user resource discovery service experiences. This study optimizes the authors' previously proposed document metadata deduplication strategy by shifting from an experience-based approach to an automated one, enhancing the automation level of the entire process while ensuring deduplication and fusion effectiveness. [**Method/Process**] Given that different document types have varying metadata items and that the same document from different sources may also differ in metadata items—both of which affect deduplication methods—this paper proposes an automated multi-source document metadata fusion model based on decision trees. The model transforms the deduplication problem into a classification problem, selects features and constructs decision trees according to feature similarity, implements metadata deduplication and fusion based on this framework, and validates the strategy's effectiveness through experiments on different document types. [**Result/Conclusion**] Results show that for five document metadata types, the deduplication strategy achieves over 99% accuracy and over 98% recall, demonstrating excellent overall performance. For fusion strategy effectiveness, the metadata quality improvement ratios for patents, dissertations, journal articles, conference papers, and books are 15.15%, 36.80%, 15.29%, 52.63%, and 15.38% respectively, all showing significant improvement.

Keywords: Multi-source metadata; Decision tree; Metadata deduplication; Metadata fusion

---

1. Introduction

In the big data environment, exponential growth of various document resources has made it difficult for traditional libraries to meet complex and evolving resource development needs, while the maturation of resource discovery systems has facilitated open integration and collaborative sharing of diverse document resources. In these systems, metadata plays a crucial role—on one hand, it helps users quickly access needed resources; on the other hand, processing massive metadata through extraction, mapping, and import effectively avoids organizational barriers caused by different sources, resources, and data structures, thereby assisting service providers in improving system management and service

levels. However, resource discovery systems also face issues such as inconsistent metadata depth due to processing costs and system adaptability, uneven quality, annotation errors, and poor interoperability between systems. Single-source metadata cannot resolve these problems. Therefore, multi-source document metadata must be fused and reorganized to comprehensively improve document metadata quality in resource discovery systems from the perspectives of accuracy and completeness, enhance metadata management and utilization, fully realize the organizational value of discovery services, and ultimately improve user experience.

To address these metadata quality issues, Lin et al. [**?**] proposed a literature metadata quality improvement model based on multi-source data fusion. Experiments showed that for journal article metadata, this strategy achieved 99.9% accuracy and 99.2% recall [**?**], demonstrating excellent overall effectiveness. However, different document types have different metadata items, requiring different deduplication methods; similarly, the same document from different sources may have different metadata items, potentially necessitating different deduplication approaches. The aforementioned journal metadata deduplication strategy relied primarily on empirical selection, resulting in limited generalizability. To address these limitations, this paper constructs an automated multi-source document metadata fusion model based on decision trees, transforming the deduplication problem into a classification problem. The model collects multi-source literature metadata, constructs features based on metadata items, calculates feature similarity for feature selection and decision tree construction, implements deduplication, forms metadata awaiting fusion, and finally generates accurate, consistent, and complete descriptions of document resources through fusion processing.

2. Related Research

As an important theme in library and information science research and practice, metadata has received extensive attention from scholars in related fields both domestically and internationally. Research most relevant to this paper includes three aspects: metadata quality evaluation indicators, metadata quality control, and metadata fusion.

Evaluating metadata quality is a prerequisite for quality control and fusion. In foreign studies, J.R. Park et al. conducted a random questionnaire survey of US metadata practitioners, revealing that completeness, accuracy, and consistency are considered the most fundamental and far-reaching indicators affecting metadata quality [**?**]. B. Stvilia et al., based on exploring the causes of information quality changes, proposed evaluation indicators including completeness, accuracy, consistency, and complexity [**?**]. T.R. Bruce et al., after refining the indicator framework proposed by B. Stvilia et al., expanded metadata quality evaluation indicators to seven aspects: completeness, accuracy, expectation satisfaction, consistency, usability, timeliness, and provenance [**?**]. Current domestic research primarily approaches metadata evaluation indicators from

two perspectives: general systems and specific systems. In general system research, scholars have summarized indicators including completeness, accuracy, accessibility, understandability, usability, compliance, timeliness, consistency, openness, and objectivity [**?**]. In specific system research, Zhang Xiaojuan et al. proposed that metadata quality evaluation indicators for government data open platforms include existence, consistency, and openness [**?**]; Dong Wei et al., considering current library development status, divided open journal metadata quality indicators into seven aspects: accuracy, completeness, timeliness, uniqueness, consistency, validity, and relevance [**?**]; Liu Jiazhen et al. argued that electronic records management metadata evaluation indicators include description level, description precision, and data currency [**?**].

Research on metadata quality control reveals that metadata quality issues mainly include insufficient cataloging standardization, accuracy, and completeness; inadequate metadata depth; metadata duplication; and deduplication errors [**?**, **?**, **?**]. To address these issues, scholars have conducted corresponding metadata quality control research covering several core components: metadata cleaning, mapping, deduplication, and phased process control. G.L. Li et al. noted that the cleaning process should focus on data quality, time efficiency, and cost control to ensure comprehensive effectiveness [**?**]. Li Huijia et al. studied semantic mapping strategies for institutional name metadata from four sources: WoS, EI, CNKI, and CSCD [**?**]. In the deduplication phase, scholars primarily focused on metadata value equality [**?**]. For phased process control, H. Manguinhas et al. introduced the UNIMARC bibliographic metadata schema and used XML format for metadata quality records to automate the quality control process [**?**]. Cao Yuezhen et al. emphasized that metadata quality control standards should be comprehensive, covering all stages from standard formulation, metadata processing, system entry, updates, inter-system interoperability, to metadata evaluation to achieve full-process control [**?**].

Quality issues in metadata structure and content in resource discovery systems lead to poor interoperability between systems, necessitating multi-source metadata fusion to resolve these problems. Wang Liya et al. utilized NLP algorithms for deduplication, normalization, and disambiguation of health medical data, and built a health big data platform [**?**]. Yan Chengxi et al. argued that metadata is a non-standardized input unit requiring conversion and composition, and their conflict redundancy detection ideas and knowledge merging methods provided references for this paper's metadata deduplication and fusion [**?**].

In summary, current research on metadata quality evaluation indicators, quality control, and metadata fusion is mature. Scholars have proposed highly similar quality evaluation indicators, primarily focusing on accuracy, consistency, and completeness, with slight variations in specific applications. Metadata quality control mainly addresses standardization, mapping, deduplication, and process control management. Metadata fusion research reveals that metadata problems include incompleteness, insufficient accuracy, unstructured data, and non-

standardization. Therefore, to address these metadata issues, this paper draws on general evaluation indicators and quality control concepts from metadata fusion research, aiming to obtain accurate, consistent, and complete metadata. The paper improves existing deduplication strategies by allowing partial equality of metadata values, improving recall while maintaining accuracy, expanding the proportion of multi-source metadata fusion, and ensuring the accuracy, consistency, and completeness of document resources as much as possible.

3. Multi-Source Document Metadata Fusion Based on Decision Trees

To obtain accurate, consistent, and complete document resources, it is necessary to acquire as many high-quality metadata descriptions as possible. Based on the analysis of metadata fusion obstacles in Lin et al. [**?**], this paper designs a multi-source document metadata fusion model based on decision trees. The model mainly includes four modules: multi-source document metadata collection, metadata preprocessing, feature selection and metadata deduplication, and metadata fusion, as shown in Figure 1 [Figure 1: see original paper]. First, the multi-source document metadata collection module collects metadata from multiple sources as the basis for deduplication and fusion. Second, the collected multi-source metadata undergoes normalization processing to facilitate subsequent deduplication and fusion. Third, features are constructed based on metadata items, feature similarities are calculated for feature selection and decision tree construction, and deduplication is implemented accordingly. Finally, the metadata fusion module fuses the metadata awaiting fusion to ultimately generate accurate, consistent, and complete metadata.

3.1 Multi-Source Document Metadata Collection To obtain accurate, consistent, and complete document resources and achieve cross-platform one-stop retrieval in resource discovery systems, metadata deduplication and fusion are necessary. Document metadata in resource discovery systems originates from multiple literature databases, such as CNKI, Wanfang Database, Baidu Wenku, Microsoft Academic Search, and Web of Science. Therefore, this paper selects different knowledge service platforms as document resource metadata source databases to collect various types of literature metadata, including patents, dissertations, journal articles, conference papers, and books. Additionally, since different document types contain different metadata items, the collected metadata items also vary accordingly.

3.2 Metadata Preprocessing Based on metadata collection, preprocessing is required to normalize metadata from various sources according to specified formats, facilitating subsequent deduplication and fusion. The preprocessing phase should follow several principles: (1) When establishing norms, design based on commonalities in cataloging content and formats across sources. Note that when two metadata items have different names but identical content, unify the names (e.g., standardize “page count” to “pages”). (2) Maintain consistent precision according to resource discovery system cataloging standards (e.g., for

time and location). For instance, in conference papers, some sources catalog the full conference location name "Xiamen, Fujian, China," while others only catalog "Xiamen." Since resource discovery system standards require only local information, retain "Xiamen" during normalization. (3) Perform fine-grained splitting of metadata items. Composite metadata items must be 拆解 to the finest granularity to maintain uniformity. For example, in Hubei University of Technology's book metadata, three items—"publication place," "publisher," and "publication time"—are merged into a single "publication distribution item." Such composite items interfere with deduplication and fusion, while specific items (e.g., publisher) play important roles in these processes. Therefore, such metadata items must be split. (4) Remove empty values. Metadata items with only names but no content must be eliminated. For example, if all "DOI" values are empty in conference papers, this item should be removed.

3.3 Feature Selection and Metadata Deduplication Service providers must determine which metadata to include in the deduplication or fusion system, requiring deduplication of normalized metadata. Previous methods relied on empirically formulated rules, resulting in low automation and efficiency. Moreover, different document types have different metadata items requiring different methods, and the same document type from different sources may also have different metadata items requiring different approaches, making original methods poorly adaptable. The essence of deduplicating two metadata records is to determine whether they describe the same document by combining various metadata items, with only a yes/no answer. Therefore, this paper transforms the deduplication problem into a classification problem, employing the decision tree algorithm from machine learning to achieve automated metadata deduplication. Features are constructed based on metadata items, similarities between features are calculated, features are selected based on similarity, and decision trees are generated. The process can be divided into several steps.

3.3.1 Feature Construction Features are constructed based on metadata items, and similarities between features are calculated. Different feature types require different similarity calculation methods, such as cosine similarity, common word ratio, Manhattan distance, and discrete values of 1, 0, -1. Since metadata items may be incomplete, data must be grouped for training, with the grouping criterion being that at least one metadata item in each pair is non-empty.

(1) For title metadata items, different metadata providers may follow different cataloging rules (e.g., not cataloging subtitles or not processing superscripts/subscripts) or make cataloging errors, leading to inconsistent title information and lengths. Therefore, drawing on the shortest matching concept, the common word ratio method is used to calculate title similarity (see Formula 1) [**?**].

$$Sim(M_a, N_b) = \frac{num_{mutual}(M_a, N_b)}{num_{shorter}(M_a, N_b)} \quad \text{Formula (1)}$$

Where M_a is title feature a in metadata information base M, N_b is title feature b in metadata information base N; $Sim(M_a, N_b)$ is the title similarity between M_a and N_b; $num_{mutual}(M_a, N_b)$ is the total number of common Chinese characters or words in M_a and N_b (without deduplication of repeated characters or words); $num_{shorter}(M_a, N_b)$ is the number of Chinese characters or words in the shorter of the two features.

Notably, before calculating title similarity, consistency of title serial numbers should be assessed. Some literature resources (e.g., books) may contain serial number information such as "Volume 1, 2, 3" or "continued," which creates strong associations between resources with highly consistent metadata values, easily causing misjudgment. Therefore, when serial numbers are inconsistent, title feature similarity is determined to be 0 to improve deduplication strategy accuracy.

(2) For other metadata item types, since only consistency (not degree of consistency) needs to be determined, discrete values are used for similarity representation in feature construction. Values 1, 0, and -1 represent feature similarity: 1 indicates identical features, 0 indicates different features, and -1 indicates one feature is empty and cannot be compared.

3.3.2 Feature Selection and Decision Tree Construction Decision trees are typically constructed using ID3, C4.5, C5.0, and CART algorithms [**?**]. Since the ID3 algorithm has low computational complexity, produces easily understandable results, and adapts well to discrete data, this paper selects this algorithm based on the characteristics of the collected literature data. The core of the ID3 algorithm lies in information gain. Therefore, information gain is defined as the amount of information a feature can bring to the classification system—that is, the degree to which information complexity is reduced under a certain condition. The greater the information amount, the more important the feature, the larger the corresponding information gain, and the greater the contribution to reducing overall system complexity [**?**]. Information gain represents the change in information before and after dataset partitioning, i.e., the degree to which the uncertainty of original dataset D is reduced after introducing feature A, as shown in Formulas (2), (3), and (4).

$$infoGain(D|A) = H(D) - H(D|A) \quad \text{Formula (2)}$$

Where $infoGain(D|A)$ is information gain, $H(D)$ is the information entropy of D, and $H(D|A)$ is the conditional entropy of D given A.

$$H(D) = -\sum_{k=1}^{K} \frac{|C_k|}{|D|} \log_2 \frac{|C_k|}{|D|} \quad \text{Formula (3)}$$

Where $H(D)$ represents information entropy, $|D|$ is the total number of data samples, K is the number of features, $|C_k|$ is the sample count for the k-th feature, and $\frac{|C_k|}{|D|}$ is the probability value of the random variable.

$$H(D|A) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|} \quad \text{Formula (4)}$$

Where $H(D|A)$ represents conditional entropy, D_i is a random variable, $\frac{|D_i|}{|D|}$ is the probability of a selected feature category, and $|D_{ik}|$ is the sample count of a certain category under D_i conditions.

Calculate the information gain after introducing each feature, and select the feature with maximum information gain as the optimal choice. Continue selecting features sequentially until categories are completely identical or no more features are available, thereby obtaining the final decision tree.

3.3.3 Metadata Deduplication via Decision Tree Different sources may produce different values for the same metadata item. Relying on a single metadata item as the deduplication condition may lead to misjudgment. Therefore, it is necessary to identify multiple metadata items about the same document and combine them for fusion; otherwise, new data errors may be introduced. For deduplicating metadata items from different sources but belonging to the same content, this paper adopts the deduplication strategy from Lin et al. [**?**]: title similarity must meet a certain threshold, while other metadata items require partial value equality (e.g., publication number, publication date, and applicant for patents). This strategy has certain fault tolerance, balancing accuracy and recall to avoid significant negative impacts on deduplication effectiveness.

3.4 Metadata Fusion After deduplication, metadata describing the same document from various sources are identified, and content fusion is performed to improve overall metadata quality and obtain accurate, consistent, and complete document resources. Based on the perspective analysis of metadata in Lin et al. [**?**], this paper employs different metadata fusion strategies for different metadata items, specifically: deduplication-based fusion, complementary fusion, rule-based fusion, and weighted voting-based fusion. The first two strategies target metadata itself and are simplest and most effective; the latter two involve quality and heterogeneity issues of source metadata. Different strategies should be adopted for different situations during fusion.

3.4.1 Deduplication-Based Fusion Strategy In most cases, metadata information from various sources is relatively complete, and metadata about the same document resource is non-empty and consistent. In such scenarios, the simplest deduplication-based fusion strategy is applied, retaining metadata from any single source. For example, for the "title" metadata item in patents, if all sources have the value "Intelligent Firefighting Robot," this value is directly used as the fused "title" metadata item.

3.4.2 Complementary Fusion Strategy While most metadata information is relatively complete, some metadata items may be missing, requiring empty value filling through complementary fusion strategy. For metadata from various sources identified as describing the same document through deduplication strategy, when only one source has a non-empty value, that value is retained. For example, for the "author" metadata item of a journal article, if only one source has a non-empty "author" item while others are empty, the non-empty information is retained as the document's "author" information. The complementary fusion strategy effectively compensates for completeness issues caused by empty values, producing significant effects despite its simple implementation.

3.4.3 Rule-Based Fusion Strategy The rule-based fusion strategy primarily addresses scenarios where metadata items from various sources for a document are all non-empty, but only one source's metadata cataloging standard complies with specifications, which is then retained. For example, for "title" metadata items from various sources, when only one source's "title" item contains a subtitle, it is used as the document's "title" information. Similarly, if the "title" metadata item contains special characters, it is split into text and special character components, retaining compliant text and special characters as the "title" information.

3.4.4 Weighted Voting-Based Fusion Strategy The weighted voting-based fusion strategy primarily addresses scenarios where metadata items from various sources for a document are all non-empty and compliant with cataloging standards, requiring fusion through weighted voting. For each metadata item, its weight is calculated according to Formula (5), and the item with the maximum weight is selected as the final result.

$$W_j = \sum_{i=1}^{k} S_{i-j} \quad \text{Formula (5)}$$

Where W_j is the weighted voting weight of metadata item j; S_{i-j} is the quality score of metadata item j from source i, derived from quantitative evaluation results in the metadata perspective analysis phase and determined based on specific circumstances.

4. Experimental Design

To comprehensively validate the model, sample data should cover various document types. CNKI and Wanfang Database are commonly used literature resource databases in domestic research, providing extensive resource coverage and comprehensive content, offering convenient knowledge access services for scholars. Therefore, this paper selects these two platforms as literature resource databases, choosing patents, dissertations, journal articles, and conference papers as metadata sources. University library collections have high academic value, and considering data availability and accessibility, book metadata is selected from Central China Normal University Library and Hubei University of Technology Library.

4.1 Data Collection The author exported 6,000 and 6,050 patent metadata records from CNKI and Wanfang Database respectively; 5,358 and 4,742 dissertation metadata records; 21,798 and 28,453 journal article metadata records; and 5,104 and 1,125 conference paper metadata records. From Central China Normal University Library and Hubei University of Technology Library, 2,488 and 2,484 book metadata records were exported. Specific metadata items are shown in Table 1 .

Table 1. Metadata Items by Document Type and Source

Document Type	CNKI	Wanfang Database
Patent	Title/Patent Name, Publication Number, Applicant, Author/Inventor, Application Date, Publication Date, Province/Country Name, Patent Category Name, Abstract, Claims, Database, Classification Number, ISSN	Title, Application/Patent Number, Publication/Announcement Number, Applicant, Inventor/Designer, Application Date, Publication/Announcement Date, Claims, Abstract

Document Type	CNKI	Wanfang Database
Dissertation	Title, Author, Keywords, Institution, Abstract, Album, Special Issue, Classification Number, Supervisor	Title, Abstract, DOI, Keywords, Author, Degree-Granting Institution, Degree Conferred, Discipline, Supervisor Name, Degree Year, Language, Classification Number, Publication Time
Journal Article	Journal Name, Title, Author, Keywords, Institution, Abstract, Volume, Issue, Year, Pages, Page Numbers	Title, Abstract, DOI, Keywords, Author, Author Institution, Journal Name, Year, Volume, Issue, Journal Column, Classification Number, Publication Time, Pages, Page Numbers
Conference Paper	Title, Author, Keywords, Author Institution, Abstract, Funding, Conference Name, Conference Time, Conference Location, Special Issue, Album, Classification Number, Page Numbers, Pages	Title, Abstract, DOI, Keywords, Author, Conference Name, Author Institution, Parent Literature, Conference Time, Conference Location, Language, Classification Code, Page Numbers

Document Type	CNKI	Wanfang Database
Book	Central China Normal University: MARC Number, Call Number, Title, Responsibility, Publisher, Publication Year, Standard Number, Document Type	Hubei University of Technology: Title, Responsibility, Publication Place, Publisher, Publication Time, Price, Physical Description, Subject, Classification Number, Abstract/Notes

The model is validated using experimental data. For deduplication strategy validation, accuracy and recall rates are used; for fusion effectiveness validation, metadata item quality improvement ratio is used. Accuracy and recall calculations are shown in Formulas (6) and (7):

$$Precision = \frac{T}{m} \times 100\% \quad \text{Formula (6)}$$

Where T is the number of records identified as duplicates by the deduplication strategy that are actually duplicates, and m is the total number of records identified as duplicates.

$$Recall = \frac{T}{T+n} \times 100\% \quad \text{Formula (7)}$$

Where n is the number of duplicate records not identified by the deduplication strategy.

4.2 Metadata Deduplication and Fusion First, preprocessing is performed on the five types of metadata, including: unifying metadata names and formats, maintaining consistent precision across metadata items, performing fine-grained splitting of metadata items, and removing empty metadata values.

Second, features are constructed using metadata items and feature similarities are calculated. Selected metadata items must have at least one non-empty value to enable similarity calculation. For title features, common word ratio method is used to obtain title similarity; other features use discrete values 1, 0, and -1 for similarity representation: 1 for identical features, 0 for different features,

and -1 for cases where one side is empty. Table 2 shows the actual metadata items selected for deduplication and fusion from various sources.

Table 2. Actual Metadata Items Selected for Deduplication and Fusion

Document Type	Deduplication Items	Fusion Items
Patent	Title/Patent Name, Publication Number, Applicant, Author/Inventor, Application Date, Publication Date	Title/Patent Name, Publication Number, Applicant, Author/Inventor, Application Date, Publication Date, Application Institution, Province/Country Name, Patent Category Name, Application/Patent Number, Abstract, Claims
Dissertation	Title, Author, Keywords, Institution, Classification Number, Supervisor	Title, Author, Keywords, Institution, Classification Number, Supervisor, Abstract, Album, Special Issue, DOI, Degree Conferred, Discipline
Journal Article	Title, Author, Keywords, Institution, Journal, Volume, Issue, Year, Pages, Page Numbers, Abstract, DOI, Journal Column, Classification Number	Same as deduplication items

Document Type	Deduplication Items	Fusion Items
Conference Paper	Title, Author, Keywords, Author Institution, Conference Name, Conference Time, Conference Location, Classification Number, Page Numbers	Title, Author, Keywords, Author Institution, Conference Name, Conference Time, Conference Location, Classification Number, Page Numbers, Abstract, Funding, Special Issue, Album, Pages, Parent Literature
Book	Title, Responsibility, Publisher, Publication Year, ISBN	Title, Responsibility, Publisher, Publication Year, ISBN, Publication Place, Price, Pages, Physical Description

First, to determine the title similarity threshold, 1,000 metadata records each for patents, journal articles, dissertations, and conference papers were extracted from CNKI and Wanfang Database, plus 1,000 book metadata records from the two library databases, totaling 5,000 metadata records as a training set. Title similarities between the two sources were calculated. When the similarity threshold was set at 0.85, the deduplication strategy for all five document types achieved over 99% accuracy and recall, demonstrating significant effectiveness. Considering all five document types, the title similarity threshold is set to 0.85.

Second, feature selection and decision tree construction are performed based on feature similarity. The greater the information gain calculated from feature similarity, the more meaningful the feature is for deduplicating metadata from various sources. Features with maximum information gain are selected as root nodes, and decision trees are constructed recursively. Figure 2 [Figure 2: see original paper] illustrates the decision tree constructed for books.

Figure 2. Decision Tree for Books

Metadata records are judged as duplicates when they meet the condition of 2/3 overlapping metadata items, generating duplicate metadata collections for fusion. The fusion process includes: (1) When metadata items from both sources are non-empty and consistent, deduplication-based fusion is applied; (2) When metadata items from both sources are non-empty but inconsistent,

and only one complies with cataloging standards, rule-based fusion retains the standard-compliant metadata; (3) When metadata items from both sources are non-empty, inconsistent, but both comply with standards, weighted voting fusion is applied, introducing other sources for voting and selecting the item with greater weight; (4) When one of the two sources' metadata items is empty, complementary fusion is applied. Finally, all fusion results are reorganized to generate accurate, consistent, and complete descriptions of document resources.

4.3 Experimental Results Evaluation This paper evaluates the multi-source metadata fusion model from two perspectives: deduplication accuracy and recall, and metadata item quality improvement ratio after fusion.

First, deduplication effectiveness is evaluated. From the deduplication results of the five metadata types, 500 records each were randomly selected, totaling 2,500 records as a test set to calculate accuracy and recall. Results are shown in Table 3 .

Table 3. Effectiveness of Deduplication Strategy by Document Type

Metric	Patent	Dissertation	Journal Article	Conference Paper	Book
Accuracy	99%+	99%+	99%+	99%+	99%+
Recall	98%+	98%+	98%+	98%+	98%+

Table 3 shows that deduplication strategies for all five document types achieve over 99% accuracy and over 98% recall, demonstrating good overall performance and validating the rationality of the deduplication strategy in the model.

Second, fusion effectiveness is evaluated through metadata item quality improvement ratio. From the fused literature metadata after deduplication, 500 records each were extracted for statistical analysis of quality improvement ratios, as shown in Table 4 .

Table 4. Effectiveness of Fusion Strategy by Document Type

Metric	Patent	Dissertation	Journal Article	Conference Paper	Book
Quality Improvement Ratio	15.15%	36.80%	15.29%	52.63%	15.38%
Quality Unchanged Ratio	84.85%	63.20%	84.71%	47.37%	84.62%
Quality Degradation Ratio	0%	0%	0%	0%	0%

Table 4 shows that metadata item quality for all document types improved to varying degrees after fusion. Conference papers showed the greatest improvement at 52.63%, because Wanfang Database had many empty values in certain metadata items (e.g., page numbers, author institutions), which were effectively supplemented through complementary fusion. Dissertations ranked second at

36.80%, primarily because Wanfang Database did not annotate keywords according to the original dissertation text, and accuracy was improved through weighted voting fusion. The other three document types showed quality improvement ratios around 15%, because their metadata items already had relatively high accuracy and completeness.

5. Conclusion

To better utilize metadata resources, promote open integration and collaborative sharing of various document resources, and optimize user resource discovery service experiences, this paper designs a decision tree-based multi-source document metadata fusion model to address metadata quality issues. The model implements deduplication and fusion, shifting from experience-based to automated approaches, expanding model applicability. Effectiveness is validated using metadata of various types from CNKI, Wanfang Database, Central China Normal University Library, and Hubei University of Technology Library. Experimental results demonstrate that the strategy achieves good fusion effectiveness for all document types, improving automation and efficiency while maintaining effectiveness. However, this study has limitations requiring future improvement: (1) Only two sources of Chinese-language literature were selected for each document type, preventing validation of weighted content fusion strategies; (2) Only Chinese literature metadata was processed, without validation for multi-language literature metadata fusion, which should be addressed in future research to enhance model applicability.

References

[1] Lin Xin, Li Xiang, Li Jing. Improving literature metadata quality in resource discovery systems based on multi-source data fusion[J]. Information Studies: Theory & Application, 2021, 44(5): 122-126, 186. [2] PARK J R, TOSAKA Y. Metadata quality control in digital repositories and collections: criteria, semantics, and mechanisms[J]. Cataloging & Classification Quarterly, 2010, 48(8): 696-715. [3] STVILIA B, TWIDALE M B, SMITH L C, et al. Information quality work organization in Wikipedia[J]. Journal of the American Society for Information Science and Technology, 2008, 59(6): 983-1001. [4] BRUCE T R, HILLMANN D I. The continuum of metadata quality: defining, expressing, exploiting[C]//HILLMANN D I, WEATHERBROOKS E L. Metadata in practice. Chicago: American Library Association, 2004: 238-256. [5] Huang Ying, Li Jianyang. Research on metadata quality assessment methods and models[J]. Library Science Research, 2013, (12): 52-56. [6] Zhai Jun, Tao Chenyang, Li Xiaotong. Research progress and implications of open government data quality assessment[J]. Library, 2018(12): 74-79. [7] Huang Gang, Yuan Man, Wu Xiuying. Research on metadata-driven data quality assessment system architecture[J]. Computer Engineering and Applications, 2013, 49(8): 114-119, 181. [8] Zhang Xiaojuan, Tan Jing. Research on metadata quality assessment of provincial government data open platforms in China[J]. E-Government, 2019(3): 58-

71. [9] Dong Wei, Zhao Jie. Research on metadata quality management of open journal resources[J]. China Science & Technology Resources Review, 2018, 50(3): 82-86. [10] Liu Jiazhen, Liao Ru. Quality control and management of electronic records management metadata[J]. Library and Information Service, 2009(6): 91-96, 102. [11] Kou Jingjing, Jia Junzhi. Comparison of Chinese retrieval performance of university library resource discovery systems[J]. Journal of the National Library of China, 2016, 25(6): 71-79. [12] LI G L, WANG J N, ZHENG Y D, et al. Crowdsourced data management: a survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(9): 2296-2319. [13] Li Huijia, Ma Jianling, Zhang Xiuxiu, et al. Research on metadata semantic mapping process: a case study of the Chinese Academy of Sciences institutional name authority control database[J]. Library Tribune, 2017, 37(12): 72-79. [14] Sun Rui, Yang Xinya, Wei Qunyi, et al. Research on key issues in literature asset metadata repository construction: a case study of Chongqing University Library[J]. Journal of Academic Libraries, 2018, 36(2): 18-24. [15] Lu Dan, Li Xin. Integration strategies for heterogeneous local chronicle metadata in digital humanities environments[J]. Library Tribune, 2019, 39(4): 158-165. [16] MANGUINHAS H, JOSE B. Quality control of metadata: a case with UNIMARC[J]. ECDL, 2006, 4172(3): 244-255. [17] Cao Yuezhen, Ma Jianling. Research progress and development trends of metadata quality control at home and abroad[J]. Library and Information, 2013(6): 101-104. [18] Wang Liya, Qiu Hang, Chen Ruoya. Health medical big data governance methods and visual presentation based on metadata traceability[J]. Chinese Journal of Health Informatics and Management, 2019, 16(6): 661-666. [19] Yan Chengxi, Fang Xiaoke. Open world perspective: a knowledge fusion framework for multi-source vocabularies MtFFO research[J]. Journal of Library Science in China, 2017, 43(4): 114-129. [20] Chu Guang, Hu Xuegang, Zhang Yuhong. Semantic-based concept drift detection algorithm for text data streams[J]. Computer Engineering, 2018, 44(2): 24-30. [21] Li Jing, Hu Qian. Automatic generation of MOOC course notes in multi-language UGC environments[J]. Information Studies: Theory & Application, 2021, 44(11): 173-179. [22] Tang Liang, Li Fei. Research on security situation prediction model for Internet of Vehicles based on decision tree[J]. Computer Science, 2021, 48(S1): 514-517. [23] Li Yongnan. Research on application of information gain decision tree in counter-terrorism intelligence analysis[J]. Information Science, 2018, 36(4): 80-84, 149. [24] Wu Peng, Xiao Weicong, Chu Rongzhen. Research on credibility of financial public opinion based on model checking[J]. Journal of the China Society for Scientific and Technical Information, 2020, 39(6): 619-629.

Author Contributions

Li Jing: Responsible for paper writing and experiment implementation; Hu Qian: Responsible for topic selection and paper revision; Li Xiang: Participated in experiment implementation and literature collection; Xiao Bing: Participated in experiment implementation and data analysis.

Research on Metadata Fusion of Multi-Source Documents Based on the Decision Tree

Li Jing, Hu Qian, Li Xiang, Xiao Bing School of Information Management of Central Normal University, Wuhan 430079

Abstract: [Purpose/significance] Constructing a multi-source document metadata fusion model will help improve the overall quality of document metadata, promote metadata management and utilization in resource discovery systems, and optimize user resource discovery service experiences. In view of the document metadata duplication judgment strategy proposed by the authors previously, this paper optimizes the strategy from experience-oriented to automated, improving the automation level of the entire process on the premise of ensuring duplication judgment and fusion effects. [Method/process] Considering that different types of documents have different metadata items, and the same document from different sources also has different metadata items, which makes duplication judgment methods different, this paper proposes an automated multi-source document metadata fusion model based on decision trees. It transforms the duplication judgment problem into a classification problem, selects features and constructs decision trees according to feature similarity, implements metadata duplication judgment and fusion on this basis, and takes different types of document resource metadata as examples to conduct experiments to verify the effectiveness of the strategy. [Result/conclusion] The results show that for five types of document metadata, the accuracy of the duplication judgment strategy reaches over 99%, and the recall rate reaches over 98%, with good overall effect. For judging the effect of the fusion strategy, the quality improvement ratios of metadata items for patents, dissertations, journal papers, conference papers, and books are 15.15%, 36.80%, 15.29%, 52.63%, and 15.38% respectively, all showing significant improvement.

Keywords: Multi-source metadata; Decision tree; Metadata duplication judgment; Metadata fusion

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