Resource Aggregation in Digital Archive Annotation Systems Based on Topic Maps: A Postprint

Authors: Zhang Yunzhong, Feng Shuangshuang

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

Abstract

[Purpose/Significance] To address the resource retrieval and navigation issues arising from the application of social tagging systems to digital archival resource organization, this study proposes a digital archival resource aggregation model based on topic maps, aiming to improve retrieval efficiency and establish orderly visual navigation.

[Method/Process] Building upon an analysis of current research on utilizing topic maps for resource aggregation in social tagging systems, this paper constructs a topic map-based resource aggregation model for the digital archives domain. It presents a systematic solution for extracting the three essential elements of digital archival resource topic maps—topic types, association relationships, and resource guidance—through social network analysis and formal concept analysis, thereby achieving resource aggregation in digital archival tagging systems.

[Results/Conclusion] Taking the "Women at War" topic in the NARA digital archival tagging system as an example, the proposed method is applied in conjunction with the Ontopia tool to implement digital archival resource aggregation for the target topic, effectively enhancing the retrieval efficiency and navigation effectiveness of digital archival resources within the system.

Full Text

Abstract

[Purpose/Significance] Aiming to address the resource retrieval and navigation problems arising from the application of social tagging systems to digital archive resource organization, this paper proposes a digital archive resource aggregation model based on topic maps to improve retrieval efficiency and establish orderly visual navigation. **[Method/Process]** Building upon an analysis

of existing research on using topic maps to achieve resource aggregation in social tagging systems, we construct a resource aggregation model for the digital archives domain based on topic maps. We present a systematic solution that employs social network analysis and formal concept analysis to extract the three key elements of digital archive resource topic maps: topic types, association relationships, and resource occurrences, thereby realizing resource aggregation in digital archive tagging systems. [**Result/Conclusion**] Taking the "Women at War" topic in the NARA digital archive tagging system as an example, we apply the proposed method combined with Ontopia tools to achieve digital archive resource aggregation for the target topic, effectively improving the retrieval efficiency and navigation effectiveness of digital archive resources in the system.

Keywords: social tagging systems; topic maps; resource aggregation; digital archives
Classification Number: G254.11
DOI: 10.13266/j.issn.0252-3116.2018.14.014

The use of social tagging systems to organize digital archive resources represents a novel approach to resource organization that has emerged in archival practice in recent years. The citizen archive tagging system of the U.S. National Archives and Records Administration (NARA) exemplifies this practice and has garnered widespread attention from archival scholars and enthusiasts. Social tagging offers clear advantages for organizing digital archive resources, including collective intelligence, timely updates, flexibility, and enhanced user experience. However, inherent deficiencies such as poor semantic standardization of tags and flat tag structures have led to retrieval and navigation problems. Scholars have reached a consensus that leveraging other knowledge organization methods to optimize tag semantics and achieve semantic-based digital archive resource aggregation is key to solving this challenge [1]. Topic maps offer a viable solution. The three core elements of topic maps—topic, association, and occurrence—enable the precise description of formal semantic relationships between topics and between topics and resources, forming an intuitive visual navigation map. Their advantages in standardization, formalization, accuracy, and visualization complement the characteristics of tags, suggesting that their combination could effectively address the deficiencies of social tagging systems for digital archives. Indeed, numerous scholars have already conducted distinctive research along these lines [2-4], attempting to establish mapping relationships between tags and topic maps from different perspectives.

This study aims to build upon existing domestic and international research by reconstructing the mapping scheme between the social tagging system triple (tags, resources, tag-resource relationships) and the three elements of topic maps. We attempt to employ quantitative tools such as social network analysis and formal concept analysis to make the mapping process more scientific, rigorous, and less subjective, thereby ensuring that the resulting digital archive resource aggregation based on topic maps can describe more precise semantics and demonstrate more accurate navigation.

1 Literature Review

Few studies have addressed digital archive tagging system resource aggregation using topic maps, but valuable insights can be drawn from research on similar problems in digital library resource aggregation and academic blog resource aggregation. The core issues in these studies focus on two questions: (1) Can topic maps clear the skies of tag clouds? This metaphor, originating from TMR2007 [5], essentially explores the feasibility of combining topic maps with social tagging systems. Such research typically examines the integration from social and cognitive perspectives. For example, D. Hendel [6] investigated the application of topic maps in social websites from social and cognitive angles, affirming the feasibility of combining topic maps with tags. Chen Ting [2] similarly affirmed this feasibility from perspectives of knowledge organization, semantic association, and technical complementarity. Scholars both domestically and internationally generally recognize the value of combining the two approaches to optimize digital resource organization. (2) How can topic maps clear the skies of tag clouds to achieve resource aggregation? Regarding how to use the combination of tags and topic maps for social tagging system organization and aggregation, scholars have employed diverse methods across various domains. K. Fujimura [7] used data mining techniques in a blog navigation system to organize and sequence large-scale tag clouds using topic maps to reveal tag relationships. Xiong Huixiang [8] and Deng Min [9] extracted topic types based on tag classification and subjectively assigned topic associations to construct topic maps in the Douban movie tagging system. Xia Lixin et al. [10] introduced a scheme for using topic maps to interconnect tags in a Fuzzy tagging system within a knowledge expert academic community. Xiang Xingbin [5] adopted a similar approach to define topic types, associations, and resource occurrences for tag resources in engineering construction, establishing a new resource organization model for tag topic maps.

In summary, existing research provides a valuable solution framework and general approaches for using topic maps to achieve information resource aggregation across various domains. However, unresolved issues remain: (1) Topic type selection typically relies on custom definitions based on tag classification and existing standards, resulting in coarse semantic granularity; (2) Determination of topic associations often depends on subjective judgment, lacking objective analysis processes and reference standards; (3) Resource occurrences are often neglected, with the aggregation process not being highlighted; and (4) Definitions of topic types and topics focus on describing external characteristics of information resources while neglecting content features. These four issues constitute the key problems this study aims to address.

2 Model Construction for Social Tagging System Resource Aggregation Based on Topic Maps

The essence of resource aggregation in social tagging systems based on topic maps is to reorganize resources originally presented in tag form using topic map structures. The key challenge can thus be abstracted as establishing a mapping from the social tagging system set {tag set, resource set, tag-resource relationship set} to the topic map {topics, associations, occurrences}. The general approach in existing research is to classify tags and map them to topic types and topics, map subjectively defined inter-tag relationships to topic associations, and map resource URI identifiers to occurrences. As mentioned in the literature review, this mainstream mapping approach has limitations. To address these, this study proposes a new mapping scheme: (1) Adopt a "cluster first, classify later" approach, replacing subjective top-down classification with bottom-up clustering to map tags to topic types and topics, making topic division more scientific with finer semantic granularity. (2) Use objective conceptual relationship analysis to extract inter-topic relationships instead of manually defined topic relationships, mapping tag relationships to topic associations to make the establishment of hierarchical and related relationships more objective. (3) Provide a detailed aggregated resource occurrence scheme to map resource sets to occurrences, enabling resources to be displayed and navigated in aggregated form.

To clearly illustrate this scheme's rationale and tasks, this study constructs a resource aggregation model for social tagging systems based on topic maps, encompassing three main modules: data processing, data analysis, and result presentation, as shown in Figure 1 [Figure 1: see original paper].

2.1 Data Processing Module

The data processing module aims to preprocess the {resource set, tag set, tag-resource relationship set} extracted from social tagging systems to lay the foundation for data analysis. Key preprocessing steps include grouping, removal, correction, and merging: (1) Grouping: This study focuses on building topic maps from the perspective of resource content features, requiring tags to be first distinguished based on whether they describe external or content features of resources. The tag set describing content features is the primary data object of interest. (2) Removal: Eliminate resources without tags and meaningless or invalid tags. (3) Correction: Fix misspelled or incorrectly written tags. (4) Merging: Consolidate abbreviations, singular/plural forms, capitalization variations, and personal/place names.

2.2 Data Analysis Module

The data analysis module utilizes specific analytical methods on the refined dataset to conduct topic and topic type analysis, association analysis, and occurrence analysis, establishing the mapping relationship from {resource set, tag

set, tag-resource relationship set} to {topic type set, topic association set, occurrence set}, thereby achieving resource aggregation in social tagging systems based on topic maps.

2.2.1 Topic and Topic Type Analysis Topics are the basic knowledge units in topic maps describing objective things in abstract form. Topics can be grouped into topic types, with a single topic potentially belonging to multiple topic types. Topic types can be extracted from both external and content features of resources. Since tags in social tagging systems describe both external and content features, selecting and extracting topics and topic types from them is essential for resource aggregation based on topic maps.

This study focuses on building topic maps from the perspective of resource content features. Using the refined "tag-resource" dataset, we adopt a "cluster first, classify later" approach. By constructing a high-frequency tag co-occurrence matrix and employing social network analysis tools to determine semantic distances between tags, we cluster the tag set into several tag groups to discover topic types, as shown in Figure 2 [Figure 2: see original paper]. The selection of high-frequency tags follows the same approach as high-frequency keyword selection in bibliometrics and will not be elaborated here. To ensure aggregation accuracy, this study uses two clustering tools—NetDraw and NodeXL—for mutual validation. In summary, selected high-frequency tags can be regarded as topics, and the resulting tag groups can be named as topic types.

2.2.2 Association and Association Type Analysis Associations are key elements revealing semantic relationships between topics and connecting related topics to form a complete semantic network. They are typically defined through expert experience, which is inevitably subjective. Therefore, this study employs formal concept analysis from applied mathematics to identify and determine semantic relationships between topics, making the process more objective. Formal concept analysis theory uses mathematical formal concepts as basic knowledge units, describes relationships between concept intensions and extensions using formal contexts, and abstracts various relationships among concepts, attributes, and instances through generalization and specialization in concept lattices, making it suitable for solving the topic relationship problem in this study.

Building on the previous step, using a selected topic type (i.e., a tag group from clustering) and its contained topics (i.e., tags within the group) as the data source, we load the data into a formal context according to the "tag-resource" binary relationship, then convert it into a concept lattice to obtain hierarchical relationships among tags. This essentially uses clustering algorithms to group resources with the same theme, transforming topics within a topic type from unordered to ordered structures. Assuming the formal context shown in Figure 3 [Figure 3: see original paper] is loaded from data contained in a certain topic type, where tag i represents the intension of a formal concept, resource j represents the extension, and "×" represents the tag-resource correspondence, it can

be converted into the concept lattice shown in Figure 3. In this lattice, nodes 1 and 2 have a genus-species relationship, which can be used to infer that topic A and topic D have a genus-species relationship. Similarly, the intersection of nodes 2 and 3 is node 4, indicating a relevance relationship between topics B and D. Thus, using formal concept analysis as a tool, we can reveal various association types including inclusion, genus-species, and relevance relationships from the perspective of resource content features.

2.2.3 Occurrence Analysis Occurrences refer to the process of linking resource entities under corresponding topics after establishing topics and their associations. Resource entities are independent of topic maps and include web pages, images, data, text, videos, and other resources that describe specific topics. These can be existing resources within social tagging systems or externally linked resources. Occurrences typically use HTML, URI, Number, Datetime, String, Image, and other occurrence types to define relationships between topic types and resource entities. This study focuses on weighted occurrence guidance for STS resources, emphasizing resource aggregation and navigation based on topic maps.

As shown in the right half of Figure 3 [Figure 3: see original paper], after establishing topic relationships, we can see that resources as extensions have hierarchical relationships. According to formal concept analysis theory, this can be interpreted as the reverse inheritance of concept extensions. This reverse inheritance can be used to describe resource weights for sorting and prioritization during resource retrieval. For example, before establishing aggregated occurrence guidance, searching for tag D retrieves resources 3, 4, 5, and 6 with equal weights. After adopting aggregated occurrence guidance, the same resource set is retrieved but with different weights—resources 3 and 5 should be prioritized over resources 4 and 6, which are generated through reverse inheritance.

2.3 Result Presentation Module

The result presentation module aims to describe and display the three types of analysis results to users using specific topic map construction tools, ultimately achieving resource aggregation in social tagging systems based on topic maps. Currently, mainstream topic map construction tools include TM4J, tinyTIM, XTM4XMLDB, and Ontopia. In this module, this study selects Ontopia, which is relatively frequently used by scholars, to "describe" the three types of analysis results for topic map construction: topic types and their topics can be described using the TopicTypes module in Ontopia; associations and association types can be created using the AssociationTypes module, covering inclusion, genus-species, relevance, and other relationships; occurrences can be created using the OccurrenceTypes module to define corresponding resource attributes, resource types, and resource links for each topic.

3 Case Study: NARA Digital Archive Resource Aggregation Scheme Based on Topic Maps

3.1 Data Acquisition and Cleaning

This study primarily uses English tag resources from the "tagging missions" section of the Citizen Archivist Dashboard in the NARA digital archives as the data source. Using the Octoparse scraper, we collected tags that users applied to 381 archives under the "Women at War" tagging mission, totaling 1,836 tags as of September 26, 2017. The collected tags were imported into Excel for manual cleaning operations including grouping, removal, correction, and merging according to the rules shown in Table 1 . After cleaning, we obtained 248 archive records with 1,695 tags.

After tag cleaning and organization, we extracted high-frequency tags following the approach used in bibliometrics for high-frequency keyword selection, as shown in Table 2 . The frequency screening rule was: first select tags with frequency $ $2 (92 tags total), with a median frequency of 4, then use tags with frequency $ $4 as high-frequency tags. Based on the high-frequency tags in Table 3, we obtained 53 tags with a total frequency of 702. Using Excel's pivottable function, we generated a 53×53 co-occurrence matrix (partial view shown in Table 3).

3.2 Data Analysis

3.2.1 Topic and Topic Type Analysis This step aims to discover topic types and their contained topics by analyzing the strength of tag relationships through clustering. To ensure clustering accuracy, this study uses both NetDraw and NodeXL for cross-validation.

(1) Importing the 53$×$53 co-occurrence matrix into NetDraw, we use the "Centrality Measures" function in the "Analysis" menu with "Degree" (describing the number of direct connections from a specific node to other nodes) as the measurement element to analyze the central position of high-frequency tags in the network and the semantic proximity between tags. The numbers in the tag co-occurrence matrix represent the co-occurrence frequency between row and column tags, with larger numbers indicating stronger associations. These associations indirectly reflect relationships among the tagged archive resources. Through clustering analysis, we can visually identify four major nodes—women, WorldWarI, WorldWarII, and posters—as the four topic types in this example, with associated high-frequency tags as their contained topics, as shown in the left side of Figure 4 [Figure 4: see original paper].

(2) Similarly, importing the tag co-occurrence matrix into NodeXL and applying its clustering function with appropriate algorithms, we cluster the tag set into tag groups, obtaining the four tag co-occurrence clusters shown on the right side of Figure 4 [Figure 4: see original paper]. The topic types are also women, WorldWarI, WorldWarII, and posters. In summary, we

identify these four key tags as the topic types for the "Women at War" tagging mission.

3.2.2 Association and Association Type Analysis This stage analyzes the association relationships and types among the obtained topic types and topics. Using the "posters" topic type and its topics as an example, we apply formal concept analysis theory. We load the topic set-resource set into a binary table, using topics as the intension of formal concepts and archive resources as the extension, with "x" representing the topic-resource binary relationship to construct a formal context. Then, using the concept lattice construction tool (conexp1.3), we convert this formal context into a corresponding concept lattice Hasse diagram for hierarchical clustering of topics, as shown in Figure 5 [Figure 5: see original paper]. Analyzing the intensions and extensions in this Hasse diagram, we identify three relationship types: genus-species, inclusion, and relevance. For example, the top-level "posters" as a topic type includes all topics, representing an inclusion relationship. The topics "flag" and "American flag" demonstrate the inheritance relationship of formal concepts, where "American flag" is a sub-topic of "flag," representing a genus-species relationship. The topics "united states army" and "women's army corps" have a relevance relationship. Other associations are analyzed similarly.

3.2.3 Occurrence Analysis This stage links resource entities under corresponding topics based on the previous analysis. Using the topics "flag" and "american flag" under the "posters" topic type as examples, according to the analysis results, the topic "american flag" should link to four archives with identifiers 515462, 514947, 513673, and 533765, while the topic "flag" should link to nine archives: 31488352, 26432783, 6788430, 533657, 535600, 515462, 514947, 513673, and 533765. According to formal concept analysis theory, the latter four archives can be considered as reverse inherited from the topic "american flag." Using this aggregated occurrence approach, searching for "flag" returns nine results, but the first five archives should be prioritized over the four obtained through reverse inheritance.

3.3 Creating Topic Maps for NARA Digital Archive Tagging System Using Ontopia

This stage uses the Ontopia topic map editor (ontopoly), browser (Omnigator), and visualization (Ontopia Navigator) tools in the OKS suite to edit, browse, and visualize topic maps, achieving resource aggregation for the digital archive tagging system.

3.3.1 Creating Topic Maps Using Ontopoly Ontopoly consists of an ontology editor and an instances editor. In this stage, we first edit the ontology content for "Women at War"—including topic types and topics, associations and association types, and occurrences—through Ontopoly's type index and

configuration pages. Then we use the instances editor to input instances for each topic, thereby creating the topic map, as shown in Figure 6 [Figure 6: see original paper].

(1) Using the TopicTypes module to create topics and topic types, we input the content feature topic types "women," "WorldWarI," "WorldWarII," "posters" and external feature topic types such as year and country/region. We configure their properties and add occurrence types like HTML, Image, and related resource links.

(2) Using the AssociationTypes module to create associations and association types, we can describe inclusion, genus-species, relevance, and other relationships. For example, in Posters, the concept "flag" has a subordinate concept "american flag," which can be edited as a "genus-species relationship" in this module.

(3) Using the OccurrenceTypes module to create resource attributes and types corresponding to topics, as shown in Table 4 . For example, in the type configuration page for the topic "flag," we can add occurrence attributes such as description, resource source, and type name, as well as resource types like HTML, URI, Number, Datetime, String, and Image.

3.3.2 Browsing Topic Maps Using Omnigator The resource aggregation and navigation results based on topic maps can be displayed through Omnigator's main interface. Omnigator [11] is a standard web browser that allows users to directly browse topics, associations, occurrences, and their links, and click through to corresponding information resources, thereby connecting internal topics and associations with information resources, as shown in Figure 7 [Figure 7: see original paper]. The browsing interface displays the "WorldWarI" topic type and topic instances in text format. Clicking on Subject Identifiers links to the web page corresponding to the "posters" tag.

3.3.3 Visualizing Topic Maps Using Navigator Topic map visualization uses a network structure to express semantic relationships between topics. Figure 8 [Figure 8: see original paper] is generated by the Ontopia Visual Navigator component, displaying the inherent and potential knowledge structure of NARA digital archive tagging resources in a network graph. Numbers on each topic indicate associated topics. For example, the number 2 on topic "flag" indicates two associated topics: the topic type "posters" and the topic "american flag." Users can select topics for further tracking queries, improving both precision and recall in retrieval. Topic maps can visually display entire topics and topic types, as well as specific associations and relationships, and even each association's linked topics.

Conclusion

Using topic maps for resource aggregation has become an effective remedy for the inherent problems of resource retrieval and navigation based on tag matching in social tagging systems. Taking NARA digital archive tagging system resource aggregation as an example, this study reconstructs the mapping scheme between social tagging system triples and topic map elements, proposes a resource aggregation model for social tagging systems based on topic maps, and employs quantitative tools such as social network analysis and formal concept analysis to more scientifically and rigorously establish the mapping from social tagging system triples to topic map elements, ensuring that the resulting digital archive resource aggregation can describe more precise semantics and demonstrate more accurate navigation.

This study has several limitations, including the small sample size due to the inability to automatically batch-obtain NARA datasets, the focus on content features at the expense of external features of archive resources, and the lack of quantitative evaluation of the digital archive resource aggregation effectiveness based on topic maps. These issues will be addressed in future research.

References

[1] RATH H. Topic maps: templates, topology, and type hierarchies [J]. IEEE signal processing magazine, 2000 (2): 45-64.
[2] Chen Ting, Hu Gaili, Chen Fuji, et al. Knowledge organization model for digital libraries in social tagging environment: from the perspective of tag topic maps [J]. Information theory and practice, 2015, 38(3): 63-70.
[3] Hu Juan, Cheng Xiufeng, Ye Guanghui. Research on knowledge organization model of academic blogs based on topic maps [J]. Library and information service, 2012, 56(24): 127-132.
[4] Xiang Xingbin. Research on construction of knowledge tag topic maps for construction enterprises [J]. Information system engineering, 2016(4): 96-97.
[5] Cleaning the skies: from tag clouds to topic maps [EB/OL]. [2017-04-29]. http://www.topicmaps.com/tm2007/lavik.pdf.
[6] HENDEL D, KUZHABEKOVA A, CHAPMAN W. Mapping global research on international higher education [J]. Research in higher education, 2015, 56(8): 861-882.
[7] FUJIMURA K, IWATA T, HOSHIDE T, et al. Geotopic model: joint modeling of user's activity area and interests for location recommendation [C]// ACM international conference on web search & data mining. New York: ACM, 2013: 375-384.
[8] Xiong Huixiang, Deng Min, Guo Siyuan. Research on construction and implementation of tag topic maps [J]. Library and information service, 2014, 58(7): 107-112.
[9] Deng Min. Research on tag semantic mining based on topic maps [D]. Wuhan: Central China Normal University, 2014.
[10] Xia Lixin, Zhang Yutao. Research on constructing knowledge expert

academic community based on topic maps [J]. Library and information service, 2009, 53(22): 103-107.
[11] Omnigator: the topic map browser [EB/OL]. [2017-05-05]. http://www.ontopia.net.

Author Contributions:
Zhang Yunzhong: Conceptualization, framework design, methodology, writing and revision of main sections;
Feng Shuangshuang: Literature collection, data collection, experimentation, initial draft writing and revision.

Abstract: [**Purpose/significance**] Aiming at the problems of resource retrieval and navigation which are caused by social tagging system used for digital archive resource organization, a digital archive resource aggregation model based on topic maps is presented in order to improve the efficiency of digital archive resource retrieval and establish an orderly visual navigation. [**Method/process**] Based on the analysis of using topic maps to realize the research status of social tagging system resource aggregation, a resource aggregation model based on topic maps in the field of digital archives is constructed, and a systematic solution to the three key elements of digital archive resource: topic types, association types and occurrence types is given, which uses social network analysis and formal concept analysis, so as to realize the resource aggregation of digital archives tagging system. [**Result/conclusion**] Taking the topic of "Women at War" in NARA digital archives tagging system as an example, we use the method proposed in this paper and combine Ontopia tools to achieve the aggregation of digital archives resources of target topic, which effectively improves the retrieval efficiency and navigation effect of digital archives resources.
Keywords: social tagging systems; topic maps; resource aggregation; digital archives

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