ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201710.00016

Organization and Exploration of Fine-grained Historical Knowledge on Contemporary China Based on Semantic Mining

Authors: Zhang, Zhixiong, Wang, Ying, Sun, Hui, Lei, Feng, Zhang, Zhixiong

Date: 2017-10-20T00:00:00+00:00

Abstract

China possesses a vast repository of historical resources pertaining to its contemporary history. A wealth of valuable knowledge remains latent within these resources and cannot be readily utilized. Addressing the urgent challenge of mining implicit semantic knowledge dispersed across extensive historical corpora and reorganizing historical knowledge and facts at a fine-grained level is essential to facilitate user exploration for research and educational purposes.

This paper proposes a method termed "Mining down, Organizing up" for the semantic representation and organization of historical knowledge on contemporary China concealed within historical encyclopedia texts. Grounded in a proposed historical ontology of contemporary China, this method extracts knowledge objects and facts from unstructured historical text items through text mining technologies, represents historical knowledge in a semantically enriched manner, and interlinks related historical knowledge objects and facts to construct a historical knowledge network of contemporary China. By mining historical facts and the historical knowledge network, the authors derive more valuable patterns from the historical knowledge that could be employed to formulate a new organizational scheme for reorganizing historical knowledge in a more dynamic fashion.

Based on this method, the authors developed a system capable of representing and organizing historical knowledge of contemporary China at a fine-grained level, supporting users in exploring historical knowledge through functions such as semantic retrieval, clustering of historical objects and facts, visual navigation, association analysis, and chronicle facts reconstruction, among others.

ChinaRxiv 中X

Full Text

Preamble

Organization and Exploration of Fine-Grained Historical Knowledge on Contemporary China Based on Semantic Mining

ZHANG Zhixiong[1], WANG Ying[1], SUN Hui[2] & LEI Feng[2]

ABSTRACT

China possesses a vast repository of historical resources on its contemporary history, within which valuable knowledge remains hidden and difficult to access. An urgent challenge is to mine the implicit semantic knowledge scattered across these resources and reorganize historical knowledge and facts in a fine-grained manner to support research and education.

This paper proposes a "Mining Down, Organizing Up" approach to semantically represent and organize historical knowledge on contemporary China embedded in historical encyclopedia texts. Grounded in a dedicated historical ontology for contemporary China, this method employs text mining technologies to extract knowledge objects and facts from unstructured historical text items, enriches the semantic representation of historical knowledge, and interlinks related historical knowledge objects and facts to construct a historical knowledge network for contemporary China. By mining historical facts and this knowledge network, the authors derive valuable patterns that enable new organizational schemes for more dynamic knowledge reorganization.

Based on this method, the authors have developed a system that represents and organizes historical knowledge of contemporary China at a fine-grained level, supporting users in exploring historical knowledge through functions such as semantic retrieval, clustering of historical objects and facts, visual navigation, association analysis, and chronicle fact reconstruction.

KEYWORDS: Knowledge Organization, Knowledge Representation, History of Contemporary China, Semantic Mining

1. Introduction

China possesses an enormous volume of historical resources on its contemporary history. With advances in digitization and networking technologies, information resources on contemporary Chinese history are growing at an accelerating pace. However, since most of these resources exist as unstructured textual data, rich semantics and substantial historical knowledge remain concealed within them, making representation, discovery, and utilization difficult. It has become an urgent challenge to mine implicit knowledge scattered across numerous historical resources and represent and organize this knowledge in a fine-grained manner for use in historical research and education.

Currently, many researchers use the term "rich semantics" to refer to implicit knowledge embedded in textual resources. Beyond historical materials, numerous documents contain rich semantics such as facts, experiences, opinions, and other information. If these rich semantics could be extracted automatically or semi-automatically from text resources, they would support a broad range of applications. For example, in the medical field, researchers have attempted to extract rich semantics from medical texts for processing, representation, and storage to enable further analysis (Kerstin Denecke 2016).

In the field of Chinese history, researchers have also applied semantic technologies to extract, represent, and organize rich semantics for knowledge discovery. The most common approach for representing and organizing historical knowledge is through ontology. Several ontologies have been developed to describe and organize historical knowledge, including the "Kuomintang-Communist Cooperation"Historical Ontology (Dong et al. 2006), the "Northeast Anti-Japanese Struggles" Historical Ontology (Wu 2012), the "Zizhi Tongjian" Historical Ontology (Peng and Song 2010), and the "Three Kingdoms" Ontology (Liao 2011). Some researchers have designed semantic platforms to represent and process historical knowledge. For instance, Prof. Dong and his team constructed a knowledge base based on semantic data from the Chinese Twenty-Four Histories and developed a Basic Historical Analysis Platform to discover implicit knowledge in historical records (Dong et al. 2014).

International work on historical knowledge organization also provides valuable inspiration. For example, Hyvönen built a historical event ontology for Finnish history and developed the semantic portal "CultureSampo," a platform for Finnish culture on the Semantic Web 2.0 (Hyvönen et al. 2007). Corda proposed a logical model of event ontology for exploring associations in history (Corda et al. 2011). Ide and Woolner outlined a model for historical ontologies that is temporally contextualized and can represent relationships among entities across different time periods (Ide and Woolner 2007).

Building on these related works, this paper proposes a "Mining Down, Organizing Up" method to organize fine-grained historical knowledge on contemporary China. Based on this method, we have developed a system that supports users in exploring historical knowledge through functions such as semantic retrieval, clustering of historical objects and facts, visual navigation, association analysis, and chronicle fact reconstruction.

2. Framework of "Mining down, Organizing up"

The "Knowledge Web of the History of the People' s Republic of China" project aims to popularize historical knowledge on contemporary China and promote historical education. One major challenge we face is employing semantic technologies to help history experts automatically or semi-automatically extract important historical knowledge from resources such as the "Dictionary of the History of the Chinese Communist Party,""Encyclopedia of the National History

of the People' s Republic of China," "Conspectus of Chinese Modern History," and "Chronicle of the People' s Republic of China." Another challenge is organizing and representing this historical knowledge so that users can discover more interesting insights when exploring the knowledge base.

To address these challenges, this paper proposes a "Mining Down, Organizing Up" method to semantically represent and organize historical knowledge on contemporary China hidden within historical encyclopedia texts. Grounded in a dedicated historical ontology for contemporary China, this approach uses text mining technologies to extract knowledge objects and facts from unstructured historical text items, represents historical knowledge in a semantically enriched manner, and interlinks related knowledge objects and facts to form a historical knowledge network for contemporary China. By mining historical facts and this knowledge network, we derive valuable patterns that enable new organizational schemes for more dynamic knowledge reorganization.

The framework is illustrated in Figure 1 [Figure 1: see original paper].

Figure 1 Framework of "Mining Down, Organizing Up"

Specifically, "Mining Down" is a deconstruction process that transforms knowledge in historical text resources into historical knowledge objects, facts, and text items that collectively form the historical knowledge network. Using text mining technology, this process automatically extracts knowledge objects from unstructured text items, annotates important sentences, and performs relation extraction to identify facts from those sentences. All extracted historical knowledge objects and facts are verified or corrected by history experts, converting unstructured text items into structured facts such as "object1-relation-object2" or "object1-property-value."

Facts about a knowledge object may be extracted from the same or different text items. Different objects can also be associated directly or indirectly through relationships (or facts) between them. Moreover, text items can be interlinked based on shared objects or facts. In this way, the historical knowledge network is constructed through associations among knowledge objects, facts, and text items. Using semantic web mining techniques such as ranking, clustering, interlinking, relation mining, and pathway analysis, more valuable patterns can be discovered from the historical knowledge, enabling fine-grained reorganization of contemporary Chinese historical knowledge.

"Organizing Up" is a construction process that uses the historical knowledge network built through "Mining Down" and new organizational schemas derived from semantic web mining to reorganize historical knowledge in more dynamic ways. Currently, historical text items on contemporary China are typically organized only by historical epochs (e.g., "1949.09-1956.09"period, "1956.10-1966.04" period) and major event categories (e.g., political events, political conventions, foreign affairs), which prevents effective discovery of hidden knowledge. Based on the constructed historical knowledge network, "Organizing Up" reorganizes historical knowledge using patterns and new organizational schemes from the

"Mining Down" process, developing a system that helps users explore historical knowledge through semantic retrieval, visual navigation, relevance analysis, and chronicle fact reconstruction.

3. Methods

Several key problems must be resolved to implement the proposed "Mining Down, Organizing Up" method: developing a foundational ontology to describe objects and relationships in contemporary Chinese historical knowledge; identifying core knowledge objects to guide extraction; extracting facts about knowledge objects from text resources; and performing semantic mining to generate new organizational schemas. This paper proposes specific methods to address these challenges.

3.1 Ontology definition

Rich semantics are embedded in historical resources on contemporary China. It is essential to determine what types of rich semantics should be extracted and disclosed to users, which requires first establishing a knowledge organization model. Therefore, we constructed a contemporary Chinese history ontology to organize extracted objects and facts.

In this ontology, we define core knowledge object types, object properties, and inter-object relations. Based on the ontology' s conceptual schema, we can outline a knowledge framework for contemporary Chinese history.

Drawing on ontology construction methods such as the Skeletal Methodology (Uschold and King 1995) and Seven Steps (Noy and McGuinness 2015), we developed the conceptual schema for contemporary Chinese history ontology with guidance from history experts. Analysis of text resources revealed that historical trajectories primarily consist of important historical events, conferences, people, and related entities. Consequently, we first defined 15 classes in the historical ontology, including Event, Conference, Person, Institution, Document, Concept, and others (see Figure 2 [Figure 2: see original paper]). Second, based on descriptions of these classes, we defined 20 datatype properties and 76 object properties to model knowledge object properties and relationships. For example, as shown in Figure 3 [Figure 3: see original paper], to represent historical event details, the ontology defines datatype properties such as label, alternative label, and literal description, and object properties such as parent event, subevent, related people, related institution, related event, occurrence time, and occurrence place. Third, we defined property restrictions—for example, parent event and subevent properties are inverse and transitive, while label and nationality properties are functional. Details of the historical ontology are available in Sun (2014).

Fig 2 Core Classes in Ontology

Fig 3 Datatype properties and object properties of Event Class

3.2 Core knowledge objects identification

To populate the ontology, we first extract metadata from historical information resources and aggregate titles of typed text items describing events, conferences, persons, or documents. Additionally, we integrate existing subject headings including person names, institution names, political parties, and geographical names. Most importantly, we obtain core historical events, conferences, and their hierarchies and associations since the founding of the People' s Republic of China, which have been manually identified and normalized by history experts. These data serve as core knowledge objects for further representation and organization.

The normalized knowledge objects include 1,685 events, 761 conferences, 3,508 persons, 2,621 institutions, 155 social groups, 107 special communities, and 1,861 hierarchical relations between events or conferences. These knowledge objects are populated into the ontology as individuals with URIs, standard labels, and alternative labels. Their relationships are represented through RDF triple statements. In the next step, we use these objects as a corpus for semantic mining.

3.3 Fact extraction

The "Mining Down" method is applied to historical text items to identify relevant facts about the core objects and populate properties and relations for corresponding ontology individuals. This process reveals knowledge hidden in text, enabling explicit expression and computational analysis. Specifically, we use text mining technology to automatically process text items, assisting history experts in establishing semantic associations between knowledge objects.

(1) Extract knowledge objects: Using a knowledge object name dictionary, we perform semantic annotation by identifying whether normal or alternative names of knowledge objects appear in text items. Additionally, we developed a named entity recognition tool to discover new knowledge objects such as time, person, institution, and conference, which are then suggested to history experts.

(2) Extract facts of knowledge objects: We detect relevant facts about knowledge objects and present candidate sentences to history experts using relation extraction technology. For example, a text item titled "The Third Plenary Session of the Eleventh Central Committee of the Communist Party of China" in the "Encyclopedia of the National History of the People' s Republic of China" describes the content of this session. The sentence "The third plenary session of the eleventh central committee, which was held in Beijing on December 18 to 22, 1978, has profound significance in the history of the Communist Party of China since its establishment" implies several facts: the holding time of "The third plenary session of the eleventh central committee" is "December 18-22, 1978," the location is "Beijing," etc. According to datatype and object properties defined in the historical ontology, we collect numerous predicate verbs such as "hold," "convene," and "take place," and create extraction rules for "conference-

time,""conference-location,"etc. Using syntactic analysis and relation extraction, facts about knowledge objects can be extracted from text items.

While automatic processing can identify potential knowledge, due to the complexity of natural language, the accuracy of text mining methods still needs improvement and results cannot be directly added to the historical ontology. All extracted information must be identified, complemented, and revised by experts based on their domain knowledge.

3.4 Knowledge network construction

Following the above steps, a knowledge network consisting of three layers—the text item layer, fact layer, and knowledge object layer—can be constructed. For example, Figure 4 [Figure 4: see original paper] illustrates the knowledge network construction process. The text item "The Third Plenary Session of the Eleventh Central Committee of the Communist Party of China" in the "Dictionary of the History of the Chinese Communist Party" reveals facts about its holding time, location, attending members, and related events. Text items with the same title in the "Encyclopedia of the National History of the People' s Republic of China" and "Chronicle of the People' s Republic of China" include not only these facts but also related concepts such as "Emancipate the Mind" and "Seek Truth from the Facts."

In the text item "The Great Historical Turning Point" from the "Conspectus of Chinese Modern History," the occurrence time, place, related conference, and related event are displayed, along with facts about related persons and the conference of "The 11th National Congress of the Communist Party of China," related persons of "The movement to criticize the 'Gang of Four' ," and so on. Thus, through a combination of text mining technology and domain knowledge from history experts, internal knowledge is discovered from texts while simultaneously constructing a complex network of historical knowledge on contemporary China.

Fig 4 Construction of knowledge network on "The Third Plenary Session of Eleventh Central Committee of the Chinese Communist Party"

3.5 Multidimensional organization

Based on the knowledge network, historical knowledge can be organized at a higher level according to relationships such as time, subclass, hierarchy, and statistics.

(1) Organization on a timeline: The time dimension provides the most direct way to show historical development processes. For instance, text items from different books can be organized by historical period, and knowledge objects with their facts can be ordered according to time classes in the ontology. Additionally, facts from the same historical period can be grouped together,

such as occurring events, held conferences, proposed policies, founded institutions, published works or documents, and presented speeches.

(2) Text item organization based on knowledge objects: Knowledge objects and facts extracted from text items provide a basis for deeper organization of those items. The same fact appearing in different source items not only verifies its accuracy but also reflects close relationships between these text items. Text items about the same knowledge object or fact can be organized together for historical biography, institutional evolution, historical data compilation, book writing, etc., providing references for research on contemporary Chinese history.

(3) Semantic organization in fact/object dimension: The contemporary Chinese history ontology effectively organizes historical knowledge and provides specific semantic representation for historical knowledge objects and facts, facilitating knowledge exploration including retrieval, association, clustering, and reorganization. On one hand, it supports fine-grained knowledge retrieval, directing users to knowledge rather than just text resources, as structured queries on ontology facts can be implemented using SPARQL. On the other hand, the same type of knowledge objects can be gathered through semantic associations, and facts about a knowledge object can be used to build a network describing that object, enabling flexible knowledge integration for various applications.

As described above, the "Organizing Up" method implements multi-dimensional display of historical knowledge at a higher level. Meanwhile, based on knowledge objects and facts, these text items can also be associated with external resources such as history books, literature, materials, web pages, or databases, enabling extension applications for contemporary Chinese historical knowledge.

4. Semantic Applications

We have applied the above method to develop a service platform providing applications including semantic retrieval, clustering of historical objects and facts, visual navigation, association analysis, and chronicle fact reconstruction.

4.1 Semantic retrieval

Unlike keyword-based search, we implemented semantic retrieval based on the historical knowledge network. When a user submits a query, the system returns knowledge objects whose preferred or alternative labels match the query term. If no match is found, it suggests similar knowledge objects or historical text items. In Figure 5 [Figure 5: see original paper], the result for a query on "Land Reform Movement" is displayed as a network showing the "Land Reform Movement" object linked to related conferences, events, documents, persons, organizations, etc. Historical objects and facts are clustered in the network, allowing users to directly obtain knowledge previously hidden in text. Additionally, users can browse source texts about the object if desired.

The platform also provides a query answering module based on the ontology. If a user asks "Who proposed 'All reactionaries are paper tigers' ," it returns the answer through a direct edge labeled "proposer"connecting the node representing the term "All reactionaries are paper tigers" to the person node "Mao Zedong," as shown in Figure 6 [Figure 6: see original paper].

Fig 6 Example of query answering

4.2 Visualization navigation

Furthermore, we implemented knowledge network visualization, where nodes represent knowledge objects and edges represent semantic relations (see Figure 7 [Figure 7: see original paper]). Users can intuitively obtain desired historical knowledge without reading full texts. The network also serves as visual navigation for discovering more knowledge by clicking on nodes, improving knowledge acquisition efficiency.

Fig 7 Fragment of visualization navigation

Figure 7 shows a fragment of visual navigation. When browsing the knowledge object "The Third Plenary Session of the Eleventh Central Committee of the Communist Party of China," we can see its attendee "Deng Xiaoping." Clicking the "Deng Xiaoping" node displays facts about him—for instance, that Deng Xiaoping proposed the concept "one country, two systems." Continuing to click "one country, two systems" shows it was proposed on "February 22, 1984." Right-clicking the edge between "one country, two systems" and "February 22, 1984" allows browsing the source text item and its context: "On February 22, 1984, when meeting guests from the United States, Deng Xiaoping explicitly proposed the concept 'one China, two systems.' In the same year on May 15, the government work report passed at the second session of the sixth National People' s Congress established 'one country, two systems' as a basic principle of national reunification." Similarly, clicking "The Government Work Report (1984)"displays its details in the network, thus implementing knowledge network navigation and browsing.

4.3 Relevance analysis

Based on the knowledge network, relevance analysis can discover potential knowledge between knowledge objects through graph traversal algorithms. For example, searching for associations between "Deng Xiaoping" and "The Third Plenary Session of the Eleventh Central Committee of the Communist Party of China" with a path length no greater than 3 produces the knowledge network shown in Figure 8 [Figure 8: see original paper]. It displays their related conferences, documents, events, persons, institutions, and links between these nodes. There exists not only a direct relationship "Deng Xiaoping → attended → The Third Plenary Session of the Eleventh Central Committee of the Communist Party of China," but also indirect links discovered through relevance analysis.

In this way, relevance analysis based on the knowledge network can uncover potential relationships between knowledge objects and enable deeper exploration of historical knowledge.

Fig 8 Example of relevance analysis

4.4 Chronicle facts reconstruction

A chronicle is a historical account of facts and events arranged along a timeline. In this study, the time class in the historical ontology can represent facts and events accurately to the month or day—for example, establishment dates of political parties, institutions, and social groups; occurrence times of events and conferences; and proposal times of principles and concepts. All this information can be used for memorabilia, chronicles, and similar applications. Figure 9 [Figure 9: see original paper] shows Chairman Mao' s historical activities in 1949. These data are computed through indirect relations between the person class and time class. For instance, it is discovered that "Mao Zedong"participated in "The First Plenary Session of the Central People' s Government Committee," which was held on October 1, 1949. Similarly, all historical activities can be generated.

Fig 9 Example of historical activity

5. Conclusions

To enable effective organization and utilization of historical knowledge on contemporary China, this paper proposes the "Mining Down, Organizing Up" method. It employs the contemporary Chinese historical ontology for semantic organization and knowledge discovery, applies text mining technology to extract important knowledge objects and facts from historical texts to form a knowledge network, and develops applications such as semantic retrieval, visual navigation, association analysis, and chronicle fact reconstruction based on this network. Studies demonstrate that the "Mining Down, Organizing Up" method enables fine-grained representation of contemporary Chinese historical knowledge and innovative knowledge organization applications based on historical knowledge objects, and can serve as a new organization and exploration method for other domains.

This study has several limitations: (1) The accuracy of recognizing knowledge objects and relevant facts from text needs improvement, particularly for national history facts, which would further reduce the workload of domain experts; (2) The association calculation method for the historical knowledge network is relatively simple and has not fully leveraged current semantic similarity calculation and graph mining methods. These are key problems to address in future research.

Acknowledgments

This article is supported by the project "Knowledge Web of the History of the People' s Republic of China" funded by the Chinese Academy of Social Sciences (Grant No. H1417) and the project "Research on Semantic Mining of Academic Resources Based on Linked Data" funded by the National Social Science Foundation of China (Grant No. 15CTQ006).

References

[Conference article] Kerstin Denecke, Yihan Deng, Thierry Declerck (2016). Extraction and Processing of Rich Semantics from Medical Texts, in Joint Proceedings of the 2th Workshop on Emotions, Modality, Sentiment Analysis and the Semantic Web and the 1st International Workshop on Extraction and Processing of Rich Semantics from Medical Texts, Heraklion, Greece, May 29.

[Journal article] Dong H., Yu C. M., Yang N., Chen L., Xu G. H., Zhang J. D., et al. (2006). Research on the ontology-based retrieval model of digital library —history domain ontology building. Journal of the China Society for Scientific and Technical Information, 2006, 25(5), 564-.

[Journal article] Wu L. J. (2012). Research on knowledge organization of characterized database based on ontology. Journal of Library Science, 2012(3), 41-43.

[Journal article] Peng W. M., & Song J. H. (2010). Research on Zizhi Tongjian historical ontology construction and application. Journal of Chinese Information Processing, 2010(2), 33-.

[Journal article] Liao Z. F. (2011). Research on domain ontology constructing and reasoning of the three kingdoms. Central China Normal University, Wuhan, China.

[Journal article] Dong H., Xu L., Wang F., & Yu S. W. (2014). Study on semantic analysis system( ) - implementation of Chinese historical records semantic analysis system. Journal of the China Society for Scientific and Technical Information, 2014, 33(2), 204-214.

[Conference article] Hyvönen E., Alm O., & Kuittinen H.(2007). Using an ontology of historical events in semantic portals for cultural heritage. In Proceedings of the Cultural Heritage on the Semantic Web Workshop at the 6th International Semantic Web Conference (ISWC 2007) (pp. 1-2). Springer.

[Conference article] Corda I., Bennett B, & Dimitrova V. (2011). A logical model of an event ontology for exploring connections in historical domains. In Proceeding of the Workshop on Detection, Representation and Exploitation of Events in Semantic Web (DeRiVE 2011), Workshop in conjunction with 10th International Semantic Web Conference (ISWC 2011) (pp. 1-10). Bonn, Germany.

[Book] Ide N. & Woolner D. (2007). Historical ontologies. In K. Ahmad, C.

Brewster, & M. Stevenson (Eds.), Words and Intelligence II: Essays in Honor of Yorick Wilks (pp. 37-152). Springer.

[Conference article] Uschold M., & King M. (1995). Towards a methodology for building ontologies. In Proceeding of the Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95 (pp. 1-13). Montreal, Canada.

[Web document] Noy N. F., & McGuinness D. L. (2015). Development 101: a guide to creating http://wenku.baidu.com/view/30fb4b956bec0975f465e2bf.html.

[Journal article]Sun H., & Lei F. (2014). Research on the contemporary Chinese history ontology building. Journal of Modern Information, 2014, 34(2):32-42.

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

Source: ChinaXiv —Machine translation. Verify with original.