Design and Construction Methodology for Knowledge Base Architecture and Its Exploratory Application in the Media Domain: Postprint

Authors: Chen Jun

Date: 2023-10-08T00:00:00+00:00

Abstract

With governments worldwide placing increasing emphasis and vigorous promotion on knowledge bases, numerous public organizations have emerged to explore knowledge base construction based on open data, with representative examples including the Linked Open Data (LOD) project and the DBPedia project, an online linked data knowledge base. In the enterprise engineering domain, major corporations such as Google, Baidu, and Sogou have also devoted themselves to research on large-scale ontology knowledge bases. As public knowledge bases become openly accessible, research on domain knowledge bases aimed at applying knowledge bases to various business domains has gradually been initiated. This paper proposes a knowledge base system design and construction methodology targeting the media application domain, constructing several major knowledge base groups including key media knowledge bases, key figure knowledge bases, key event knowledge bases, business keyword knowledge bases, and business knowledge encyclopedia bases, introduces in detail several key technologies for knowledge base construction, and provides a focused elaboration on application scenarios of knowledge bases in the media domain.

Full Text

A Design and Construction Method for Knowledge Base Systems and Its Application Exploration in the Media Domain

Abstract: With governments worldwide placing increasing emphasis on and providing strong support for knowledge bases, numerous public organizations have explored knowledge base construction based on open data. Representative

examples include the Linked Open Data (LOD) project and the online linked data knowledge base DBPedia. In the enterprise engineering domain, companies such as Google, Baidu, and Sogou have also devoted significant resources to research on large-scale ontological knowledge bases. As public knowledge bases become more open, research on domain-specific knowledge bases aimed at applying them to different business fields has gradually emerged. This paper proposes a design and construction method for a knowledge base system targeted at media applications, building several major knowledge base groups including key media knowledge bases, key figure knowledge bases, key event knowledge bases, business keyword knowledge bases, and business knowledge encyclopedia bases. The paper details several key technologies for knowledge base construction and elaborates on application scenarios of knowledge bases in the media domain.

Keywords: knowledge base; knowledge graph; ontology-based knowledge representation; knowledge extraction; knowledge annotation; knowledge base application

Author: Chen Jun

1. Research Status and Development Trends

With governments worldwide placing increasing emphasis on and providing strong support for knowledge bases, numerous public organizations have explored knowledge base construction based on open data. Among them, the Linked Open Data (LOD) project is particularly representative, using RDF to publish various open datasets on the Web. By establishing RDF links between data items from different sources and interconnecting the semantic links used by different ontological knowledge bases, LOD achieves maximum global knowledge sharing. Additionally, as an online linked data knowledge base project, DBPedia extracts structured data from Wikipedia entries to provide more accurate and direct Wikipedia searches, creating connections between other datasets and Wikipedia to deliver large-scale world knowledge across languages and domains.

In the enterprise engineering domain, Google, Baidu, and Sogou have also actively engaged in research on large-scale ontological knowledge bases. Notable examples include Google' s "Knowledge Graph," Baidu' s "Zhixin," and Sogou' s "Zhilifang." These systems integrate massive amounts of fragmented Internet information, aggregating search results centered around keywords into knowledge clusters. They re-optimize search results through recalculation, presenting users with the most core information.

As public knowledge bases become more open, research on domain-specific knowledge bases aimed at applying them to different business fields has gradually emerged. For instance, case-based reasoning knowledge base systems extract and organize knowledge from relevant cases, enabling them to recommend similar solutions and reference content for user-input problems. Ontology-based domain-specific knowledge base systems perform digital semantic processing of

thematic business materials, classifying and annotating them according to ontological principles to achieve knowledge integration, sharing, discovery, and innovation within specific business domains.

2. Overall Knowledge Base Architecture

[Figure 1: see original paper] Overall Architecture Diagram of the Knowledge Base System

The infrastructure layer primarily provides various required computing resources, storage resources, and network resources, upon which big data foundational applications are built. By offering diverse storage forms such as relational databases, document knowledge storage databases, message queues, and caching, this layer enables different types of data to be stored according to their characteristics and business requirements, thereby satisfying system real-time demands and distributed response architecture needs.

The data resource layer mainly provides standardized storage functions for various data resources related to upper-layer functions from a business perspective, while also offering unified storage for system data such as message queues and caching resources.

The key technologies layer provides the core supporting technology systems required for implementation, offering key technologies including knowledge description and acquisition, knowledge graphs, and knowledge base construction with analysis and assessment.

The system function layer provides data analysis and display functions for business personnel, as well as human-computer interaction interfaces for annotators. This layer constructs five major knowledge bases: key media knowledge base, key figure knowledge base, key event knowledge base, business keyword knowledge base, and business knowledge encyclopedia base. Each knowledge base implements unified knowledge description methods, classification and organization systems, and evaluation indicator systems to maximize compatibility with existing knowledge bases and functional modules. Each knowledge base features functions for knowledge extraction, annotation, evaluation, and maintenance, while establishing an evaluation system for annotators.

Based on these five major knowledge bases, relevant analysis and assessment functions are constructed, including knowledge association and inference, as well as business statistics and analysis capabilities. For important events, the system performs analyses such as geographic distribution, time cycles, figure distribution, pattern mining, and trend prediction.

3. Key Technologies for Knowledge Base Construction

Key technologies for knowledge base construction primarily include knowledge representation, knowledge acquisition, knowledge graphs, knowledge persistence,

and knowledge evaluation.

3.1 Knowledge Schema Construction

Knowledge representation forms the foundation for knowledge acquisition and application. Currently, the most commonly used method is ontology-based knowledge representation. An ontology abstracts the essence of entities within a domain, emphasizing associations between entities and expressing these associations through various knowledge representation elements. These elements, also known as primitives, mainly include: (1) concepts; (2) attributes; (3) relations; (4) functions; (5) axioms; and (6) instances. Overall, the purpose of constructing ontologies is to achieve a certain degree of knowledge sharing and reuse: (1) Ontological analysis clarifies the structure of domain knowledge, thereby laying a solid foundation for knowledge representation. Ontologies can be reused to avoid redundant domain knowledge analysis. (2) Unified terminology and concepts make knowledge sharing possible.

Based on the different data types of knowledge sources, knowledge is classified to form five categories of knowledge bases: important media, important figures, important events, business keywords, and business knowledge encyclopedia. Each category can be further subdivided. The knowledge tree method is employed to organize knowledge. At each level, knowledge nodes maintain consistent granularity with adjacent nodes. Higher levels have larger granularity, while lower levels have finer granularity. The system establishes corresponding tree-structured knowledge systems for each knowledge category based on user business experience. Users can edit these systems by adding or deleting nodes and resetting node names.

3.2 Knowledge Evaluation System Construction

The knowledge evaluation system is a mechanism for assessing the quality of existing knowledge, evaluating from three aspects: completeness, validity, and relevance. Knowledge completeness is directly calculated by the system based on the fill rate of knowledge entry attributes, using the calculation rule: weight $\times$ score. Business personnel jointly discuss and determine the weight values and scores for different categories of knowledge attributes, based on which a completeness percentage is given for each knowledge entry. Knowledge validity is obtained through system-business personnel interaction. When business personnel review a knowledge entry, they can evaluate its usefulness by clicking "useful" or "useless" buttons. The system displays real-time distribution statistics of usefulness ratings for each knowledge entry. Relevance is similar to usefulness, with business personnel evaluating whether a knowledge entry is business-related and assigning a relevance value. When multiple people evaluate the same knowledge entry' s relevance, the average value is displayed.

3.3 Knowledge Acquisition

Knowledge acquisition includes knowledge extraction, knowledge annotation, and knowledge maintenance. The process of building a knowledge base involves extracting knowledge from structured and unstructured data resources. Structured knowledge acquisition refers to parsing data in specific formats, such as structured database records, HTML, XML, and other semi-structured data containing tags, to obtain multiple knowledge entities and their detailed attributes, as well as the association relationships between knowledge entities.

Unstructured knowledge acquisition refers to extracting entities and relationships mentioned in documents, element keywords, and document summaries from imported text materials. Through automatic identification and extraction of content categories, this information is stored in different knowledge entry collections. Supported formats include TXT, Word, Excel, PDF, and various other forms.

The system supports manual annotation and maintenance of knowledge base entries, with knowledge annotation adoptable through crowdsourcing. Annotators can select an entity via right-click to annotate it, with annotated knowledge being interconnected across multiple data sources. If the entity to be annotated already exists in the knowledge base, intelligent prompt completion is provided to save annotation time, improve annotation efficiency, and ensure annotation consistency. Simultaneously, for each knowledge entity, associated entities with direct relationships are displayed in visual form, supporting visual editing of the entity' s associated entities and relationships.

The system supports collaborative work by multiple users on knowledge entities. After a user modifies entity attributes, if the database version is inconsistent with the version before user modification upon submission, the system alerts the user to potential conflicts. The user must obtain the new version and make modifications and submissions based on it to maintain consistency.

3.4 Knowledge Graph

Knowledge graphs research methods for knowledge association, association, and inference to achieve application modes such as knowledge inference and assessment. Knowledge association analysis mines and displays association relationships between knowledge base entities, establishing associative relationships between discrete knowledge nodes in the form of network graphs. When a node in the association graph is clicked, detailed information about that node is displayed. The system' s knowledge association supports not only connections between knowledge entities of the same category but also cross-channel associations between key figures, historical events, business keywords, business knowledge encyclopedia, and other entities.

Knowledge association is proposed to improve knowledge retrieval efficiency for business personnel during knowledge base usage. Currently, most information

retrieval employs full-text retrieval technology, with retrieval strategies based on statistical patterns of keyword frequency. Association-based retrieval recommends knowledge entries with similar semantics to user search content, providing users with alternative options.

Knowledge inference derives new, unknown knowledge from existing knowledge in the knowledge association graph to improve knowledge completeness and expand knowledge coverage, applicable to business scenarios such as similar-type knowledge search and relationship prediction.

3.5 Knowledge Persistence Technology

The purpose of knowledge persistence technology is to persistently store constructed knowledge bases. Currently, data in knowledge graphs primarily uses semantic XML document specifications and structured databases for persistent storage. However, these storage methods cannot achieve rapid knowledge queries within linear time during large-scale knowledge subgraph queries. To accelerate query speed, existing query algorithms generally employ graph indexing technology. However, knowledge graphs have large data scales, and building graph indexes for them consumes substantial time and space overhead, making it difficult for users to quickly obtain satisfactory query results. To address these characteristics, we adopt a graph structure-based storage solution for knowledge persistence, achieving fast and efficient knowledge graph storage and querying. Based on distributed graph data processing platforms, we employ new knowledge graph query models, algorithms, and computing platforms to persist knowledge from three aspects: knowledge graph query models, distributed query algorithms, and distributed query execution optimization.

4. Application Exploration of Knowledge Bases in the Media Domain

Based on the aforementioned knowledge base systems—including key media knowledge base, key figure knowledge base, key event knowledge base, business keyword knowledge base, and business knowledge encyclopedia base—the system can provide not only direct knowledge retrieval and recommendation but also various rich analytical application functions such as knowledge association and inference, heuristic search, personalized recommendation, in-depth topic planning, in-depth event analysis, trend prediction, machine reading, and machine writing. These can be applied to various news production scenarios.

4.1 Heuristic Search

For editors or information analysts, the information they want to search for is often not clearly defined. Therefore, they typically set a general analysis goal, perform initial screening from massive information, and then adjust keywords based on preliminary results to search for more precise content. During this process, through associative relationships between domain knowledge elements,

knowledge association can recommend relevant knowledge, helping users gradually perform information association analysis and deep mining from point to surface. We call this exploration-based search heuristic search. For example, when searching for "aircraft accidents," the system can recommend associated information such as historical aircraft accident tracking over the years, engines, aerospace, related manufacturing companies, and related financial stock information. Through exploration of big data based on business knowledge associations, users can obtain broader analysis perspectives, thereby mining higher value-added information and enhancing the service value of in-depth reporting products across social life, politics, industry, finance, and other domains.

4.2 In-Depth Topic Planning Analysis

When editors plan topics for one or a group of subjects, simply recommending content with similar descriptions is often insufficient. Users hope to discover entirely new angles for topics. Providing knowledge association and recommendations through interconnections between business domain knowledge offers greater value for planning in-depth reports, data journalism, and think tank consulting. For example, for smog reporting, if analysis can extend beyond smog itself to explore China' s energy consumption structure, industrial structure and layout, and further expand to the impact of domestic macro-control policies over the years and even overseas energy futures market trading situations, the breadth and depth of such reporting content analysis will be greatly enhanced. This provides comprehensiveness and innovation not found in simple homogeneous reporting, thereby significantly improving the professional level and public influence of media reporting products.

4.3 In-Depth Event Analysis

Utilizing a multi-dimensional tagging system for the media industry, the system performs multi-dimensional knowledge indexing on massive news events, enabling thematic aggregation of news content with co-reference relationships. Based on knowledge-driven approaches, various dimensions of in-depth analysis are conducted, including event location, occurrence time, event subject, related subjects, event homology relationships, causal relationships, spatiotemporal relationships, original reporting media, and relevant policies and regulations. The system tracks the daily evolution of sub-topics during event development and analyzes viewpoints expressed by important domestic and foreign figures, media outlets, and institutions regarding the event.

Trend prediction provides information on potentially important future events and their probabilities. According to specific business requirements, this can include future event prediction, keyword popularity trend prediction, and sensitive event information prediction. Future event prediction displays events that may occur within a specified future timeframe along with related information and occurrence probabilities. Keyword popularity trends show the change patterns of keywords related to the event within a specified timeframe. Sensitive

event prediction provides information on potentially sensitive events that may occur in the future. Users can customize timeframes to predict trends within specified periods.

4.4 Machine Reading

Machine reading refers to the automation of processes previously requiring human reading comprehension. Common task forms for machine reading currently include synthetic question answering, entity completion, and answer candidate prediction. Synthetic question answering involves business personnel pre-constructing corpora formed by simple facts and corresponding questions. Machines then read and comprehend article content and perform reasoning to derive correct answers. Entity completion involves machines reading and understanding corpora, after which questions are posed—often sentences with entity words removed from the article. The machine' s process of answering questions involves predicting the removed entity words in the question sentences. Answer candidate prediction involves machines, based on articles, corresponding questions, and candidate answers, understanding and reasoning to predict correct answers from candidate options. By establishing standardized entity tags and constructing knowledge graphs and domain knowledge bases, these machine reading functions can be well supported.

4.5 Machine Writing

Machine writing represents an automated trend in content production—a process of algorithm-based content generation and editing. Computers can automatically generate content in specific formats by selecting combinable content from existing candidate material libraries based on given topics, acquiring data, analyzing data, and extracting viewpoints.

During the data acquisition and analysis phases, knowledge bases can provide data related to specific topics obtained by machines and information about content mentioned in materials that is related to knowledge entries in the knowledge base, supporting preliminary data support in the machine writing process. Simultaneously, knowledge bases can enrich writing results based on known historical knowledge. During the viewpoint extraction process, knowledge entry tags in knowledge bases can also provide foundational data support for viewpoint extraction, improving the effectiveness of extracting important viewpoints from data.

References: [1] Chen Xinlei, Jia Yantao, Wang Yuanzhuo, et al. A Multidimensional Quantitative Evaluation Method for Open Knowledge Base Construction Technology[J]. Computer Science, 2017(12). [2] Yang Yuji, Xu Bin, Hu Jiawei, et al. An Accurate and Efficient Domain Knowledge Graph Construction Method[J]. Journal of Software, 2018(10). [3] Mao Hui. Knowledge Base-Based Question Answering System[J]. Modern Computer (Professional Edition), 2019(8). [3] Guan Saiping, Jin Xiaolong, Jia Yantao, et al. Research Progress on

Knowledge Reasoning for Knowledge Graphs[J]. Journal of Software, 2018(10).

(Author Affiliation: Xinhua News Agency Technology Bureau)

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