Postprint: Research on the Construction and Application of Vertical Knowledge Graphs

Authors: Ruan Tong, Wang Mengjie, Wang Haofen, Hu Fanghuai

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

Abstract

[Purpose/Significance] In recent years, knowledge graph technology has garnered widespread attention from both academia and industry. This paper proposes a data-driven incremental knowledge graph construction methodology, offering a novel approach for building vertical knowledge graphs. Additionally, application demonstrations of vertical knowledge graphs are provided through three case studies. [Method/Process] Initially, a formal definition of knowledge graphs is presented. Subsequently, a data-driven incremental knowledge graph construction method is proposed, with emphasis on investigating the intricacies and challenges involved in constructing the data graph component of vertical knowledge graphs. Leveraging this methodology, this paper constructs knowledge graphs for Traditional Chinese Medicine, marine science, and enterprise domains. [Results/Conclusion] The successful construction of the aforementioned vertical knowledge graphs validates the feasibility of the proposed method, while their respective domain-specific applications illustrate the extensive applicability of knowledge graph technology.

Full Text

Preamble

Research on the Construction and Application of Vertical Knowledge Graphs

Ruan Tong[1], Wang Mengjie[2], Wang Haofen[1], Hu Fanghuai[3]

Abstract: [Purpose/Significance]

In recent years, knowledge graph technology has garnered widespread attention from both academia and industry. This paper proposes a data-driven incremental knowledge graph construction method, offering a new approach for building vertical knowledge graphs. Additionally, it provides application demonstrations

for vertical knowledge graphs through three case studies. [Method/Process] We first present a formal definition of knowledge graphs, then propose a data-driven incremental construction method, focusing on the details and challenges of constructing the data graph component of vertical knowledge graphs. Based on this method, we have built Traditional Chinese Medicine knowledge graphs, marine knowledge graphs, and enterprise knowledge graphs. [Result/Conclusion] The construction of these vertical knowledge graphs confirms the feasibility of our proposed method, and their respective vertical applications demonstrate the broad utility of knowledge graphs.

Keywords: knowledge acquisition, knowledge fusion, semantic search, prescription assistance, relation discovery
Classification Number: TP391
Citation Format: Ruan Tong, Wang Mengjie, Wang Haofen, et al. Research on the Construction and Application of Vertical Knowledge Graphs [J/OL]. Knowledge Management Forum, 2016, 1(3): 226-234 [citation date]. http://www.kmf.ac.cn/paperView?id=43.

1 Introduction

Since the concept of the Semantic Web was proposed, vast amounts of Linked Open Data (LOD) and User-generated Content (UGC) have been published on the Internet, transforming it from a document web comprising only hyperlinks between web pages into a data web containing rich descriptions of entities and their relationships. In this context, Google introduced the concept of the "knowledge graph" in 2012 to improve search engine performance [1]: a semantic network that describes objective entities, concepts, and their associative relationships in the real world.

Based on application domains, we categorize knowledge graphs into general knowledge graphs and vertical (or industry-specific) knowledge graphs. General knowledge graphs are not domain-specific and can be analogized as "structured encyclopedic knowledge," containing extensive commonsense knowledge with an emphasis on breadth. Representative large-scale general knowledge graphs include YAGO [2], DBpedia [3], Freebase [4], and NELL [5], while Chinese general knowledge graphs include Zhishi.me [6] and SSCO [7]. Vertical knowledge graphs, by contrast, are domain-specific, built upon industry data, and emphasize depth of knowledge. They can be viewed as industry knowledge bases built on semantic technologies, with potential users being industry professionals.

In general knowledge graph construction, relatively mature technologies and products already exist, such as commercial knowledge graphs released by major search engine companies like Google's Knowledge Graph, Baidu's "Zhixin," and Sogou's "ZhiliFang." However, in vertical knowledge graph construction, existing approaches often rely on manual construction, lacking a unified methodology. Therefore, this paper focuses on vertical knowledge graphs by first providing a formal definition, then proposing a data-driven incremental construction

method that examines the details and challenges of knowledge acquisition and fusion from multiple data source types.

2 Formal Definition of Knowledge Graphs

There is no essential difference between general and vertical knowledge graphs, so this paper defines both uniformly. As shown in Figure 1 [Figure 1: see original paper], a knowledge graph G consists of a schema graph Gs, a data graph Gd, and the relationship R between them, i.e., G = <Gs, Gd, R>. The schema graph Gs = <Ns, Es>, where Ns represents the set of class nodes and Es represents the set of property edges. Classes (nodes) in Gs correspond to concepts in the knowledge graph, while properties (edges) represent semantic relationships between concepts, including standard RDFS [8] properties like rdfs:subClassOf and rdfs:equivalentClass, as well as user-defined properties such as employer. Similarly, in the data graph Gd = <Nd, Ed>, the node set includes instance nodes and literal nodes, while edges in Ed connect two nodes to represent a triple fact, such as <Bill_{Gates}, almaMater, Harvard_{University}>. Here, instances are entities representing objectively existing objects recognizable by computers, while literals often serve as attribute values for instances. The relationship R between Gs and Gd is constituted by rdf:type, representing the relationship between instances in the data graph and their corresponding concepts.

Knowledge graphs offer several technical advantages. First, they allow easy modification of data schemas with excellent dynamic extensibility, enabling incremental schema design during construction. Second, their semantic interoperability characteristics and "linked data" principles facilitate integration of data from different sources. Additionally, knowledge graphs support existing standards such as RDFS, OWL [9], and SPARQL [10], which can gradually be required from content providers. Finally, knowledge graphs explicitly express relationships between entities, enabling applications like semantic search and automatic question answering.

3 Related Work

In knowledge graph construction, substantial work has accumulated on general knowledge graphs. Early approaches relied primarily on manual construction, resulting in knowledge graphs like WordNet [11] and ResearchCyc [12]. Subsequently, many knowledge graphs were built based on Wikipedia, such as YAGO and DBpedia. However, due to different extraction targets, their knowledge richness varies [13]. DBpedia extracts all information and statistics from Wikipedia infoboxes, while YAGO only extracts custom properties from Wikipedia and integrates data using WordNet, achieving higher accuracy but lower knowledge richness than DBpedia. Unlike these tools, Zhishi.me uses Chinese Wikipedia and additionally incorporates two popular Chinese encyclopedia sites: Hudong Baike and Baidu Baike.

Recently, open-domain knowledge extraction projects like KnowItAll [14] and NELL have gained attention, using incremental iterative methods to learn high-quality triples from vast web data to build knowledge graphs. However, due to different application scopes, vertical and general knowledge graphs employ different construction methods. General knowledge graphs typically use bottom-up approaches, which facilitate discovery of new knowledge graph patterns. Vertical knowledge graphs, emphasizing hierarchical knowledge structures, usually require pre-constructed schema graphs. Since general knowledge graph construction methods are unsuitable for vertical knowledge graphs, and existing high-quality vertical knowledge graphs often rely on manual construction, this paper proposes a data-driven incremental knowledge graph construction method to provide a general approach for automatic vertical knowledge graph construction.

4 Construction of Vertical Knowledge Graphs

4.1 Overview

Since vertical knowledge graphs emphasize knowledge depth and overall hierarchical structure, construction typically combines top-down and bottom-up approaches. The top-down approach involves pre-building the schema graph of a vertical knowledge graph using ontology editors or manual methods, then constructing the data graph. The bottom-up approach involves using multiple extraction techniques to obtain entities, properties, and relations from knowledge sources during data graph construction, then merging high-confidence extraction results into the knowledge graph.

As shown in Figure 1, knowledge graph G consists of schema graph Gs, data graph Gd, and their relationship R. Assuming the schema graph Gs of a vertical knowledge graph has already been constructed, this paper adopts a bottom-up approach to explain the process of building the data graph Gd and relationship R from data sources. As illustrated in Figure 2 [Figure 2: see original paper], knowledge sources are primarily divided into structured knowledge, semi-structured knowledge, and unstructured knowledge. Structured knowledge includes large amounts of linked open data and domain knowledge stored in relational databases. Semi-structured knowledge includes infoboxes from encyclopedia sites like Wikipedia, Hudong Baike, and Baidu Baike, as well as numerous tables and lists from vertical domain sites. Unstructured knowledge refers to the vast amount of plain text content in web data, which offers the broadest knowledge coverage but poses the greatest extraction difficulty.

Starting from knowledge sources, knowledge graphs are built mainly through two steps: knowledge acquisition and knowledge fusion. Based on the characteristics of knowledge graphs themselves, we can continuously enrich the constructed knowledge graph using an incremental iterative approach. This construction process is called data-driven incremental knowledge graph construction.

4.2 Knowledge Acquisition

The knowledge acquisition stage needs to extract entities, synonym relationships, and "property-value" relationships from knowledge sources to build the data graph Gd, while also extracting entity types to construct relationship R. Due to numerous knowledge sources and data overlap between them, the challenge lies in adopting appropriate extraction methods for different knowledge source types and fully utilizing data redundancy.

Structured and semi-structured knowledge, having explicit structures and fixed formats, are relatively easy to extract, while unstructured textual knowledge is more difficult. As shown in Figure 3 [Figure 3: see original paper], for relational database data among structured knowledge, we can use D2R (Relational Database to RDF) mapping methods to transform it into linked data in the knowledge graph. For semi-structured knowledge like infoboxes and tables in encyclopedia data, we use wrapper-based extraction methods. The authors propose a multi-strategy learning approach for knowledge acquisition [15]. Multi-strategy learning leverages redundant information across different knowledge sources, using easily extractable information to assist in extracting difficult information. Wrappers are information extraction methods designed for data sources with specific structures. We extract these two types of knowledge and add the results to a seed set.

For unstructured plain text knowledge, we employ an incremental iterative extraction approach combining distant supervision [16] and pattern-based methods. Distant supervision is based on the assumption that "if two entities have a certain relationship, any sentence containing that entity pair is likely to express the same relationship." This method uses known entity relationship pairs to automatically annotate text. Here, the seed set can be used to automatically annotate text data, which then automatically generates high-quality patterns. These patterns are applied to text to learn new knowledge, which is added to the seed set. This process iterates continuously until no new knowledge can be learned.

4.3 Knowledge Fusion

The knowledge acquisition stage only extracts entities, properties, and relationships from different knowledge source types, forming isolated extraction graphs. To create a complete knowledge graph, these extraction results must be integrated through knowledge fusion. During knowledge fusion, various data conflict issues must be resolved, including one phrase corresponding to multiple entities, inconsistent entity property names, missing entity properties, inconsistent entity property values, and one-to-many mappings of entity property values. The knowledge fusion stage primarily performs entity matching and schema alignment.

Entity matching aims to discover entities with different identifiers that represent the same real-world object and merge them into a single entity object with

a global unique identifier. Clustering methods are commonly used for entity matching, with the key being to define appropriate similarity measures. These similarity measures often reference the following entity features: (1) string similarity: entities with identical descriptions may represent the same entity; (2) property similarity: entities with identical property-value relationships may represent the same object; (3) structural similarity: entities with identical neighboring entities may point to the same object.

Schema alignment primarily involves integration of entity properties and property values. For entity property integration, features to consider include property synonyms, entity types at both ends of the property, and patterns corresponding to the property during extraction. When data conflicts occur during fusion of data from different knowledge sources, factors such as knowledge source reliability and frequency of different information across knowledge sources can also be considered. The authors propose merging knowledge cards from search engines [17], providing an online knowledge fusion approach. This solution first proposes a probability-based entity scoring algorithm to find the most relevant Wikipedia entry for a knowledge card, thereby merging different knowledge cards representing the same entity. Then, it uses the mapping relationship between Wikipedia infoboxes and DBpedia ontology as training data to design four-dimensional features for training a property alignment model. Finally, similarity thresholds are used for deduplication and merging of property values to form value clusters.

5 Case Studies of Vertical Knowledge Graphs

Using the data-driven incremental knowledge graph construction method, we have built Traditional Chinese Medicine knowledge graphs, marine knowledge graphs, and enterprise knowledge graphs. The following sections elaborate on the construction process and specific applications of these three vertical knowledge graphs to demonstrate the effectiveness of our method and the broad applicability of vertical knowledge graphs.

5.1 Traditional Chinese Medicine Knowledge Graph

The Traditional Chinese Medicine (TCM) field has accumulated substantial professional knowledge classification information, which we use to build the schema graph of the TCM knowledge graph. Currently, we primarily construct disease libraries, syndrome libraries, and other sub-library schema graphs based on the Classification and Codes of TCM Diseases and Syndromes (national standard), clinical guidelines provided by the China Association of Chinese Medicine, and the drug database from Shuguang Hospital affiliated with Shanghai University of Traditional Chinese Medicine.

For constructing the data graph of the TCM knowledge graph, we use D2R mapping to extract drug information from Shuguang Hospital's relational database, construct wrappers for Microsoft Office software to extract disease and pre-

scription information from national standards like the "1998 Edition Syndrome Classification Standard" and clinical knowledge bases stored in Microsoft Word format at Shuguang Hospital, and iteratively learn plain text knowledge from encyclopedias and TCM websites using a combination of pattern-based and distant supervision methods. Due to data extraction from multiple sources, duplicates and conflicts exist between different data sources. We address data conflicts by scoring data source credibility and ranking data items based on data origin and frequency across sources.

The resulting TCM knowledge graph mainly includes disease, syndrome, symptom, herbal medicine, and prescription libraries. Based on this TCM knowledge graph, we can perform natural language question answering related to TCM. Additionally, using the Drools inference engine [18], we can provide TCM prescription assistance.

TCM prescriptions are divided into basic prescriptions and empirical prescriptions. Basic prescriptions are determined by diseases and syndromes, while empirical prescriptions require determination based on patient symptoms. After a doctor diagnoses a patient's disease, syndrome, and symptoms, the TCM knowledge graph can recommend prescriptions through reasoning. As shown in Figure 4 [Figure 4: see original paper], the TCM knowledge graph stores basic and empirical prescriptions for the "Liver Qi Stagnation Syndrome." The Drools inference engine transforms these prescriptions into a set of rules. When a patient presents with hypochondriac pain and is diagnosed with Liver Qi Stagnation Syndrome, the basic prescription can be applied according to the rules. However, for patients also experiencing "bitter taste and dry mouth," the rules require removing Chuanxiong and adding Moutan Cortex.

5.2 Marine Knowledge Graph

The marine knowledge graph primarily includes fish knowledge, marine economic knowledge, and island knowledge. Marine economic knowledge was collected by domain experts and stored in Microsoft Word documents, which we transformed into marine knowledge sub-graphs using Microsoft Word wrappers. Island knowledge originates from relational databases provided by the Zhoushan Marine Digital Library, which we converted using the D2R mapping tool D2RQ [19] to form island knowledge sub-graphs.

Fish knowledge has numerous data sources, including three major Chinese encyclopedia sites, the Taiwan Fish Database (fishdb [20]), the FishBase world fish classification hierarchy [21], industry sites like Xinshipu (recipe website), and textual data such as the "Chinese Food Composition Table" (2002 edition). To construct the schema graph of the fish knowledge sub-graph, we use HTML wrappers to extract concepts and hypernym-hyponym relationships from fishdb and FishBase, and extract concept attributes from encyclopedia pages. We use multi-strategy learning methods to iteratively extract synonym relationships from these data sources. For the data graph, we extract fish instances from

fishdb and FishBase and obtain instance property values using multiple methods. For example, we use HTML wrappers to obtain values for the "fish cuisine" property from Xinshipu, and use patterns to extract values for the "nutritional components" property from the "Chinese Food Composition Table" (2002 edition).

After review by marine knowledge experts and resolution of data conflicts, the marine knowledge graph currently contains over 30,000 named fish species globally with more than 20 properties, providing knowledge services including marine knowledge visualization, semantic retrieval, and marine knowledge recommendation. The marine knowledge graph offers three visual retrieval modes: wheel view, tree view, and detail view, focusing respectively on displaying semantic relationships between entities, the hierarchical structure of the marine knowledge graph, and detailed properties of entities and concepts. Additionally, its semantic search service provides direct answers to users' natural language questions while displaying entity knowledge cards and related entities, and returns literature search results combined with library resources. As shown in Figure 5 [Figure 5: see original paper], the question "distribution of small yellow croaker" is parsed to extract the entity "small yellow croaker" and the semantic "ecosystem distribution of small yellow croaker," based on which the system returns semantic search results, the knowledge card for "small yellow croaker" and related entities, along with relevant literature resources.

5.3 Enterprise Knowledge Graph

The enterprise knowledge graph integrates data from 30 million enterprises, along with patent data and bidding data from the internet, including executive information changes and corporate merger and acquisition information.

The enterprise knowledge graph can provide actual controller queries and relationship discovery functions. The actual controller query function identifies the natural person with the largest shareholding in an enterprise. Since personal control can be exercised through direct investment or through enterprises controlled by the individual, the algorithm is implemented based on graph traversal. Users input an enterprise, and the system returns its actual controller. The relationship discovery function can identify indirect relationships between companies or individuals. Figure 6 [Figure 6: see original paper] shows the relationship between "Aluminum Corporation of China Limited" and "CITIC Securities Company Limited," where arrows represent investment relationships. The graph reveals that the investors of these two companies jointly invested in Enterprise B.

First, domain experts constructed the schema graph of the industry knowledge graph, including top-level concepts such as person, company, stock, patent, investment, and bidding. Then, using D2R tools, we transformed enterprise information from relational databases provided by enterprises into RDF (Resource Definition Framework) data, forming the basic enterprise knowledge

graph. However, this initial enterprise knowledge was relatively simple and required supplementation from other data sources. We added patent and bidding information by crawling text announcements from the Chinese Government Procurement Network and China Patent Information Network, then used heuristic information-defined patterns to extract enterprise bidding and patent information. Subsequently, we further supplemented enterprise information based on encyclopedias and news.

6 Conclusion

This paper provides a formal definition of knowledge graphs and describes in detail a data-driven incremental knowledge graph construction method. Using this method, we constructed Traditional Chinese Medicine knowledge graphs, marine knowledge graphs, and enterprise knowledge graphs, and developed related applications. The construction of these three vertical knowledge graphs demonstrates the feasibility of our proposed method and the advantages of knowledge graphs in knowledge fusion, while their applications reflect the value of knowledge graphs in different domains.

References

[1] SINGHAL A. Introducing the knowledge graph: things, not strings[EB/OL]. [2012-05-16]. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html.
[2] BIEGA J, KUZEY E, SUCHANEK F M. Inside YAGO2s: a transparent information extraction architecture[C]//Proceedings of the 22nd international conference on World Wide Web companion. Rio de Janeiro: International World Wide Web Conferences Steering Committee, 2013: 325-328.
[3] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia-A crystallization point for the Web of data[J]. Web Semantics: science, services and agents on the World Wide Web, 2009, 7(3): 154-165.
[4] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of the 2008 ACM SIGMOD international conference on management of data. Vancouver: ACM, 2008: 1247-1250.
[5] CARLSON A, BETTERIDGE J, KISIEL B, et al. Toward an architecture for never-ending language learning[C]//Proceedings of the twenty-fourth AAAI Conference on artificial intelligence. Atlanta: AAAI Press, 2010: 3.
[6] NIU X, SUN X, WANG H, et al. Zhishi.me-weaving chinese linking open data[M]. The Semantic Web–ISWC 2011. Berlin: Springer, 2011: 205-220.
[7] HU F, SHAO Z, RUAN T. Self-supervised Chinese ontology learning from online encyclopedias[J]. The scientific world journal, 2014, 2(11):1-13.
[8] BRICKLEY D, GUHA R V. RDF vocabulary description language 1.0: RDF schema[EB/OL]. [2003-12-15]. https://www.w3.org/TR/2003/PR-rdf-schema-20031215/.
[9] BECHHOFER S, VAN HARMELEN F, HENDLEr J, et al. OWL web ontol-

ogy language reference, 2004[EB/OL]. [2004-02-10]. https://www.w3.org/TR/owl-ref/.
[10] PRUDHOMMEAUX E, SEABORNE A. SPARQL query language for RDF[EB/OL]. [2008-01-15]. https://www.w3.org/TR/rdf-sparql-query/.
[11] MILLER G A. WordNet: a lexical database for English[J]. Communications of the ACM, 1995, 38(11): 39-41.
[12] RAMACHANDRAN D, REAGAN P, GOOLSBEY K. First-orderized researchcyc: expressivity and efficiency in a common-sense ontology[C]//AAAI workshop on contexts and ontologies: theory, practice and applications. Pittsburgh: AAAI Press, 2005: 25-34.
[13] RUAN T, DONG X, WANG H, et al. Evaluating and comparing web-scale extracted knowledge bases in Chinese and English[C]//Joint international semantic technology conference. Yichang: Springer International Publishing, 2015: 167-184.
[14] SCHMITZ M, BART R, SODERLAND S, et al. Open language learning for information extraction[C]//Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning. Jeju Island: Association for Computational Linguistics, 2012: 523-534.
[15] RUAN T, LIN Y, WANG H, et al. A multi-strategy learning approach to competitor identification[M]//Semantic Technology. Berlin: Springer International Publishing, 2014: 197-212.
[16] MINTZ M, BILLS S, SNOW R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP: Volume 2. Singapore: Association for Computational Linguistics, 2009: 1003-1011.
[17] WANG H, FANG Z, ZHANG L, et al. Effective online knowledge graph fusion[M]. The Semantic Web-ISWC 2015. Berlin: Springer International Publishing, 2015: 143-157.
[18] PROCTOR M. Drools: a rule engine for complex event processing[M]//Applications of graph transformations with industrial relevance. Berlin: Springer, 2011: 2.
[19] CYGANIAK R, BIZER C, GARBERS J, et al. The d2rq mapping language[EB/OL]. [2007-08-24]. http://d2rq.org/.
[20] SHAO K T. Taiwan fish database[EB/OL]. [2014-11-10]. http://fishdb.sinica.edu.tw.
[21] FROESE R, PAULY D. Fishbase[EB/OL]. [2012-06-15]. http://www.fishbase.org.

Research on the Construction and Application of Vertical Knowledge Graphs

Ruan Tong[1], Wang Mengjie[2], Wang Haofen[1], Hu Fanghuai[3]

[1]School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237
[2]Department of Computer Science & Engineering, East China University of Sci-

ence and Technology, Shanghai 200237
[3]Shanghai Hi Knowledge Technology Co. Ltd., Shanghai 200433

Abstract: [Purpose/significance] In recent years, knowledge graphs have gained widespread attention from both academia and industry. This paper proposes a data-driven incremental knowledge graph construction method, offering a new approach for building vertical knowledge graphs. Additionally, through three case studies, it provides application demonstrations of vertical knowledge graphs. [Method/process] We first present a formal definition of knowledge graphs, then propose a data-driven incremental construction method, focusing on the details and challenges of constructing the data graph component of vertical knowledge graphs. Based on this method, we have built Traditional Chinese Medicine knowledge graphs, marine knowledge graphs, and enterprise knowledge graphs. [Result/conclusion] The construction of these vertical knowledge graphs confirms the feasibility of our method, and their respective vertical applications demonstrate the broad utility of knowledge graphs.

Keywords: knowledge acquisition, knowledge fusion, semantic search, prescription assistance, relation discovery

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